Evaluating Fairness and Generalizability in Models Predicting On-Time Graduation from College Applications

Stephen Hutt
University of Colorado Boulder
594 UCB, Boulder,
Colorado, 80309, USA
stephen.hutt@colorado.edu

Margo Gardner
University of Colorado Boulder
594 UCB, Boulder,
Colorado, 80309, USA
margo.gardner@gmail.com

Angela L. Duckworth
Character Lab & University of Pennsylvania
3401 Market St. Philadelphia,
PA 19104 USA
aduckworth@characterlab.org

Sidney K. D'Mello
University of Colorado Boulder
594 UCB, Boulder,
Colorado, 80309, USA
sidney.dmello@colorado.edu

ABSTRACT

We explore generalizability and fairness across sociodemographic groups for predicting on-time college graduation using a national dataset of 41,359 college applications. Our features include socio-demographics, institutional graduation rates, academic achievement, standardized test scores, engagement in extracurricular activities, and work experiences. We identify five latent classes based on available sociodemographic data and train Random Forest classifiers to successfully predict 4-year graduation. When individually trained and tested on each class using a split-half validation method, we achieved AUROCs between 0.629 and 0.694. We then evaluate how a model trained on the entire dataset performs on each latent class by performing a slicing analysis, finding a 6 to 10 percent improvement in AUROCs compared to the individual-class models. We explore fairness of our model by extending the slicing analysis to consider Absolute Between ROC Area (ABROCA), finding similar values for each of our latent classes. We contemplate how our results might be used to avoid perpetuating biases inherent in college application data.

Keywords

college success, college applications, generalizability, fairness, slicing analysis, National Student Clearinghouse, Common App

1. INTRODUCTION

In 2016, the Obama administration issued a report urging data scientists to explore "how technologies can deliberately or inadvertently perpetuate, exacerbate, or mask discrimination." [6]. To this point, in recent years, machine learning has come to influence a range of real-world activities, such as detecting credit fraud, financial investing, advertising, and, of course, education. These methods make complex, sometimes life changing, decisions based upon training data, often without considering if the training data is biased. This is a critical omission because training data takes advantage of past events, which may be unfairly biased against certain subpopulations, such as those of a particular race, gender, or sexual orientation. Given that the past data may be biased, machine learnt models further perpetuate or even exacerbate these biases. The resultant models can be described as 'unfair' because they treat different subpopulations differently. For example, Amazon recently had to decommission their recruitment AI as it favored male applicants for technical jobs [15], ostensibly because in the ten years of hiring data used to train the model, the company had predominantly hired more men than women in technical roles. This biased training data caused the model to negatively evaluate resumes that alluded to the applicant being female (e.g. containing the phrase "women's chess club captain" or having attended a women's college).

In the context of education, there has been considerable interest in predicting a range of educational outcomes, such as affect, learning style, likelihood of dropping out of a course, and whether a student will succeed on an upcoming test [18, 46]. In this paper, we examine generalizability and fairness in the specific case of predicting college success. We use the CommonApp-NSC dataset, a 6-year longitudinal de-identified dataset of college application data and graduation outcomes. We first identify five subgroups of applicants based on sociodemographics and then train Random Forest classifiers to predict on-time college graduation from application data. We explore how the models generalize across groups and how fair the models are to each of the groups.

1.1 Background

A college degree offers a wide variety of personal, academic, and economic benefits [3]. For example, 2015 median earnings for young U.S. adults (25-34 years of age) with a Bachelor's degree were 64% higher than those who had only completed high school, a consistent pattern over the past 15 years [39]. In addition to economic gains for the student, college completion also correlates with economic gains for the nation as a whole [11].

However, based on the latest data, only 40% of first-time, full-time U.S. students graduated with a Bachelor's degree within four years [41, 52] (60% graduated within six years). Moreover, the academic

Stephen Hutt, Margo Gardner, Angela L. Duckworth and Sidney D'Mello "Evaluating Fairness and Generalizability in Models Predicting On-Time Graduation from College Applications" In: *Proceedings of The 12th International Conference on Educational Data Mining (EDM 2019)*, Collin F. Lynch, Agathe Merceron, Michel Desmarais, & Roger Nkambou (eds.) 2019, pp. 79 - 88

achievement gap separating students by race/ethnicity in K12 persists in college. Only 21% of Black and 30% of Hispanic students graduated within 4 years, compared to 44% White and 48% Asian/Pacific Islander students [52].

These numbers suggest that there is considerable room for improvement overall and especially for closing the achievement gap. More to the point of the present focus, they introduce a substantial potential for bias in any machine learning model which uses demographics to predict 4-year college graduation outcomes since these models might inherently predict lower graduation for an underrepresented minority based on historic rates irrespective of their abilities.

In addition to demographics, socioeconomic (SES) factors have been reliably linked to college success [16, 17, 19, 63]. An early 1964 study showed that SES factors such as family income, parental occupation, and parental education levels had a significant effect on college retention and graduation [22]. Other work has reported similar links between SES and college success [1, 34, 60], particularly with respect to the relationship between SES and ethnicity [48]. This signals another potential for bias. Indeed, recent work in machine learning (and even in the popular press) has called for an examination of how SES is used in models [14, 67], especially when making crucial decisions about a person's future.

Beyond SES, academic achievement such as high school GPA and standardized test scores have also been shown to be predictive of college success. In large scale studies [53, 66], including a landmark study with 150,000 students, both standardized test scores and high school GPA predicted college success. However, both of these measures has also been (negatively) linked to SES [54, 68], suggesting potential bias in a more indirect way.

Beyond sociodemographics and cognitive ability, Goodwin and Hein [28] hypothesize that the "X-Factors" such as a can-do attitude, self-discipline, and good study habits are also important for college success. This view aligns with Duckworth [16], Dweck [19], Walton [63], and others who argue that non-cognitive factors such as grit, self-control, mindset, and social belonging [16, 17, 19, 63] are critically important for college success after accounting for sociodemographics and cognitive ability.

One complicating factor is that these non-cognitive traits are difficult to accurately measure. Therefore, admissions counselors must rely on self-reports or informer reports (such as from teachers), which have a number of known biases (see [17] for a review). To address this, there is an interest in more objective measurement approaches. One relevant proxy measure of non-cognitive traits is sustained engagement in extracurriculars during high-school. The rationale for the predictive value of extracurriculars is that they provide a context for the development and demonstration of key non-cognitive characteristics (e.g., initiative [37], identity [21, 37], competence, confidence, and character [9]) linked to academic success. However, extracurriculars might also be inherently biased, in that SES influences the amount and types of available extracurricular opportunities [31, 38]. There is some evidence that work experiences might provide similar benefits as extracurriculars provided youth do not work too much (see [65]). However, low SES students might be the ones more likely to work [61, 64], suggesting that work experiences might also be a biased proxy measure.

In previous work with the CommonApp-NSC data set [33], we trained models that could successfully predict four year graduation using sociodemographics, cognitive ability, and non-cognitive factors. However, we did not consider how our models generalized between subpopulations or if a population was being treated unfairly. We address this issue here by exploring how the models perform across different sociodemographic groups and evaluate the fairness of our classification methods.

1.2 Related Work

Because the field of generalization is vast [59], we focus specifically on generalization across sociodemographic groups in the context of education. We then go on to discuss fairness for a model – that is whether predicted outcomes for a particular group are consistently negative. Available techniques fall into two groups: (1) methods for evaluating if an existing model is fair; and (2) methods for developing a fair model. Both approaches are discussed here.

In machine learning, cross-validation is performed to improve the likelihood of a model generalizing to new instances, or in educational data mining, to new students (i.e., students not in the training set). However, what about generalizability beyond the student to groups of students? The results might not be so promising. For example, a review by Blanchard [7] indicated that work in intelligent tutoring systems (ITS) and artificial intelligence in education (AIED) overwhelmingly samples from White, educated, industrialized, rich, and democratic (the so-called WEIRD) countries. Blanchard notes that cognitive factors differ greatly across cultural contexts and stresses the importance of expanding research in ITS and AIED to non-WEIRD countries.

Baker and Gowda [4] showed that generalizability is not even guaranteed within communities in the United States of America. Using data from a diverse group of students interacting with an ITS, they found that behaviors that were predictive of disengagement significantly differed between urban, suburban, and rural students. This was further supported in [42], where models trained on urban or suburban students generalized to each other but not to rural students. In contrast, Samei et al, [49] reported that their models of effective classroom discourse generalized across urban and non-urban classrooms. Bosch et al [8] also have had success with generalizability. In work on detecting affect while students interacted with an educational physics game, they found that video-based affect detection generalized across ethnicity (as perceived by human observers) and gender.

Each of the aforementioned approaches explore generalizability by training models on one group and testing on another, a simple yet effective method of validation. However, this approach does not apply when it is infeasible to build models for each sub population, for example, when training data is limited. In this case, we must instead consider how a model trained on all the data performs to individual subgroups of interest. In slicing analysis [50], predictive models are trained on the entire data set and evaluated by "slicing" along subpopulations of interest (such as race, or ethnicity). This allows a researcher to explore if a model is only successful for a certain group (e.g. if a model trained on all data is only accurate for white students).

Gardner et al. [26] recently presented a metric to evaluate fairness within slices. They propose Absolute-Between-ROC Area (ABROCA) for quantifying how a predictive model's performance varies across different student subgroups. This metric evaluates whether a model privileges (provides more accurate classification) or disparately impacts (provides less accurate classification) a subgroup by comparing the group's ROC curve to the ROC curve of a baseline group. In a study analyzing MOOC dropout rate, they show a significant difference in privilege given to males versus

females in machine learnt models across a variety of feature sets and across a classification techniques.

Another method for evaluating models is Individual Fairness [20], which states that in order for a model to be considered fair it must yield similar predictions to similar individuals. The success of this evaluation method depends upon how similarity is defined, a challenging task when the number of predictors is large as in any complex prediction problem.

An alternative to evaluating fairness post-hoc is to design fair models from the ground up. *Fairness through unawareness* posits that a model is fair if it does not include any protected (potentially biased) variables (e.g. [25, 36]). However, this approach ignores the fact that protected variables such as ethnicity may be encoded (via correlation or similar) with variables not initially considered to be protected, such as participation in extracurricular activity or standardized test scores. [27, 32].

Alternatively. Kusner et al. [36] have introduced the idea of Counterfactual Fairness**,** which requires an understanding of causality among predictors (see [36, 45]). Whereas this method has been successful in datasets with a limited number of variables, understanding causality in a complex dataset presents many challenges and may render this approach unfeasible.

1.3 Contributions of Current Study

The present study is novel in multiple respects. We build upon work predicting on-time Bachelor's graduation solely from information contained in college applications. We derive 143 variables from each application and train models on 41,359 instances. Our sample includes students from all 50 U.S. states as well as international students, yielding many options for exploring generalizability and fairness.

We first cluster the sociodemographic data by identifying latent classes of students within the dataset. From available sociodemographic variables (e.g., race/ethnicity, parent education, parents' marital status, and English language learner status), we identify five distinct latent classes of applicants for further investigation. Specifically, we examine the accuracy of models trained and tested on the same class. We then use a slicing analysis to investigate how a model trained on all the data performs for each of the classes.

It should be noted that our complex, real-world dataset does not easily lend itself to current methods for designing fair models. Decades of research have shown that SES is a predictor of college success (as cited above), so a fairness through unawareness approach would require ignoring an important predictor. Likewise, counterfactual fairness requires understanding causality in the dataset, a challenge given that we are working with 143 variables. We instead evaluate the fairness of models trained with traditional machine learning approaches using ABROCA as the pertinent metric.

2. DATASET

2.1 CommonApp-NSC Data[1]

The Common App [55] is a nonprofit organization that hosts a portal where high school students can complete and submit applications to nearly 700 colleges. The Common App streamlines the admissions processes by enabling students to complete one "common" application that can be submitted to multiple colleges across the country. Whilst individual colleges may have their own supplemental applications (e.g. additional essays), the core application remains the same.

The Common App has three parts. The student section includes information on sociodemographics, future college plans, family history, academic history, standardized test scores, honors received (for academic, sporting, or other pursuits), extracurricular activities, work history, and disciplinary history. Students also submit a personal essay, but these are not available to us due to privacy concerns. A separate evaluation consists of teacher ratings of the student across several dimensions, ranging from "quality of writing" to "reaction to setbacks". Finally, the secondary school report contains information on the student's high school (e.g., percent of graduation class enrolling in college), the student's academic performance (e.g., class rank, GPA), and evaluations from the student's guidance counselor (e.g., ratings of academic achievements, difficulty of courses, and personal qualities).

The National Student Clearinghouse (NSC) is a nonprofit organization created in connection with the financial aid lending industry that gathers enrollment data for student borrowers. The NSC data tracks the following information for each student on a per-semester basis: college name, college type (2/4 year; private or public), enrollment (none, full, part-time), major, and graduation status (degree, and major).

Both organizations have merged, de-identified, and shared the data with us, which we prepared for statistical analyses. The Common App contains individual applications from 413,675 students who completed the 2008 application for admission in the 2009 school year. We successfully matched 362,205 of these applications to 2015 NSC records. From this subset, we removed 50,894 students who enrolled in college prior to 2008 and an additional 3 students due to data integrity issues, leaving 311,308 students.

To account for institutional effects on the probability of graduation, we obtained 4- and 6-year graduation rates from the National Center for Educational Statistics (NCES) [47] for the institution students first enrolled in (i.e., their first entry in the NSC). We used 2012 graduation rates to avoid including students from our 2009 student cohort who would be on-track to graduate with a 4-year degree in 2013. We obtained institution graduation data for 89% of the students, resulting in a reduced sample of 278,201 students.

We also obtained information on students' high school environments (e.g. demographics of the school) from the NCES data [47], using the 2007-2008 school year to avoid direct overlap with our student cohort.

Our data only included applications/reports that were completed online, as there was a paper option in 2008. Of the 278,201 students, only 41,359 had a corresponding teacher evaluation and secondary school report, which contained critical GPA scores as entered by guidance counselors. We presume that a majority of the missing cases were submitted on paper; they were therefore not available to us. Previous work investigated the importance of GPA in predicting college success [33] and found that it did not significantly boost prediction after accounting for the other features. Here, we with this subset for consistency, but do not consider GPA.

[1] In what follows, we provide an abridged description of the dataset, which was originally published in [33].

2.2 Encoding the Application

We extracted 143 features from the application, including auxiliary sources (e.g., NCES data), which we grouped these into the following categories:

Personal and family, 48 features. Features in this category focus primarily on sociodemographics (e.g., ethnicity, sex, number of parents who went to college, etc.).

Academics and standardized tests, 38 features. This category encodes information from the 'academics' (e.g., did a student intend to graduate from high-school on time) and 'standardized tests' (e.g., SAT scores) section of the student application along with data about the high school environment from the NCES (e.g., teacher-student ratio).

Activities and work experience, 45 features. Students enter information for up to seven extracurricular activities, including the type of the activity, the time commitment, and the school years in which they participated in the activity. In addition, students can enter up to three work experiences, from which we derived features such as number of jobs and hours per week at each job.

Honors, 10 features. Students describe academic and sporting honors received during their high school career. For each honor, we encode the type of honor, the level of the honor (school, state, national or international), and the grade when it was received.

Institutional graduation rates, 2 features. These are the 4- and 6-year graduation rate of the colleges in which students first enrolled.

Our sample of 41,359 students represented all 50 states and included some international students. The students represented 5,678 secondary schools and were enrolled in 1,238 post-secondary institutions. Forty-four percent (44%) graduated within four years of enrollment; this rate aligns with national norms [41].

2.3 Student demographics

Our sample was majority female (56%). Student age was unavailable due to data de-identification, which eliminated birth dates. With regard to ethnicity, 54% of students identified as Caucasian, 8% as African American, 8% as Hispanic, 8% as Asian American, 5% as Asian Indian, 4% as Mexican American, 1% as Native American/Alaskan, and 5% as other ethnicities (students could select multiple ethnicities as well as decline to answer). In terms of home life and education, 96% had two living parents, 77% of students reported living with both parents, 68% had two parents who attended college, and 16% had one parent who attended college. For secondary education, 68% of students reported attending a public high school, 14% a religious high school, 16% an independent high school, 2% a charter school, and 1% were home schooled. The subset of 41,359 students was representative of the full sample of 311,308 students, differing only with respect to the number of parents who attended college [33].

3. LATENT CLASS ANALYSIS

We used latent class analysis (LCA) to identify five clusters of students based on individual sociodemographic characteristics (race/ethnicity, parent education, parents' marital status, and English language learner status), the race/ethnic composition of students' high schools (% African American, % Latino, % White, and % Asian American), and whether the school was Title I eligible (a school is eligible if it has high concentration of low income students [23]). We selected these characteristics because they not only paint a relatively comprehensive portrait of socioeconomic status, but also have demonstrated associations with college success (see Introduction). Specifically, White and Asian American students, students of college educated parents, students with married parents, and students who speak English as a first language have higher on-time graduation rates than are African American and Latino students [58], first generation college students [57], English language learners [35], and students with single parents [44, 51]. Likewise, high schools with large percentages of low-income and minority students, when compared to predominantly White higher-income high schools, often have lower rates of college matriculation and completion [29]. Although often used in generalizability studies (e.g., [8, 24]), we did not include gender in these models as it does not relate to SES and ethnicity (see Discussion).

We used the entire sample size, in this case, all students who attended a public high school (N= 216,133) for the latent class analysis in order to obtain the most representative clusters. We used complex mixture models with a maximum likelihood estimator in MPlus 7 [40] to identify our latent class structure. Standard errors were adjusted to account for the clustering of students within high schools. An initial two-class solution yielded AIC and BIC values of 385,195.512 and 385,493.737, respectively. Subsequently, we tested solutions with up to six classes. Although each increase in the number of classes resulted in notable improvements in model fit (see Table 1), the magnitude of these improvements diminished with increasing model complexity. The selection of the final five-class solution (see Table 2) balanced model fit against pragmatism. Specifically, the six-class solution fit the data somewhat better than the five-class solution, but two of the six classes had very similar profiles (i.e., profiles similar to class 3 in Table 2). Each class in the five-class solution, on the other hand, had a distinct profile as described below.

Table 1. Model fit by number of latent classes

#	AIC	Incremental reduction in AIC	BIC	Incremental reduction in BIC
2	385195.51	--	385493.74	--
3	196766.61	188428.90	197198.52	188295.21
4	16127.75	180638.87	16693.35	180505.18
5	-103474.54	119602.29	-102775.26	119468.60
6	-180815.24	77340.70	-179982.27	77207.01

Of the 41,359 students analyzed here, only 28,122 were included in the LCA analysis since we only focused on those who attended a public high school. Class 1 contains a plurality of Black students with a sizable white minority (20%) in their high schools. The majority of students are native English speakers, approximately half of students are first generation college students and approximately half of students are children of unmarried parents. This reflects an average SES. Class 2 contains a plurality of White students, but other groups are represented in their high schools (51% white). The majority are native English speakers with married, college-educated parents. Students in this class typically attend a non-Title I eligible, diverse high school where approximately half the students are white with moderate representation across other ethnic/race groups. Thus, this class can be categorized as predominantly white students, high SES students in diverse high schools. Class 3 is similar to Class 2, except with a higher majority of white students and students typically attending a primarily white high school. These students are also high-SES.

Table 2. Five Class Solution: profiles across classification variables by latent class N= 28,122

			Proportion of class that is…									Average high school race/ethnic proportions			
Latent class/ label	**N**	**Proportion of sample**	**White**	**Black**	**Latino**	**Asian**	**Other race/ ethnicity**	**First generation college student**	**Child of married parents**	**English Language Learner**	**Attending Title I eligible high school**	**White**	**Black**	**Latino**	**Asian**
1 (Black, mid-SES)	1,745	0.06	0.26	0.51	0.10	0.08	0.10	0.52	0.54	0.14	0.40	0.20	0.61	0.12	0.03
2 (White/diverse, high SES)	5,670	0.20	0.47	0.10	0.13	0.19	0.15	0.33	0.75	0.18	0.25	0.51	0.15	0.18	0.11
3 (White, high SES)	16,959	0.60	0.69	0.02	0.04	0.07	0.20	0.23	0.79	0.05	0.20	0.85	0.04	0.04	0.03
4 (Asian, high SES)	1,051	0.04	0.17	0.02	0.05	0.67	0.10	0.29	0.87	0.52	0.20	0.28	0.05	0.13	0.52
5 (Latino, low SES)	2,697	0.10	0.17	0.11	0.54	0.13	0.08	0.70	0.62	0.45	0.77	0.13	0.12	0.65	0.06

*Note. Proportions of students of varying race/ethnic groups in each class do not sum to 1 due to missing data. N is less than the full sample as we only included students from public high schools

Class 3 is the largest of the classes representing 60% of the students. Class 4 is the smallest of the classes, containing 4% of students, the majority of students are Asian, and a sizable number (52%) of students are English language learners. The majority of students in class 4 have college educated; married parents and attend a non-Title I eligible high school, which suggests high-SES. Finally, in class 5, there is a plurality of Latino students, many (45%) of whom are English language learners. The majority of students have married parents who did not attend college. Most students in this class attended majority Latino, Title I eligible high schools, suggesting low-SES.

4. MACHINE LEARNING

We used the scikit-learn library [43] for machine learning. We focused on Random Forests because previous work that considered logistic regression, naive Bayes, decision tree (using the scikit-learn CART-like algorithm), and gradient-boosted decision trees [33]. found that Random Forest was consistently the best performing approach.

Hyperparameters for the random forest classifier [10, 30], were tuned on the training set using the cross-validated grid search method provided by scikit-learn [43]. Specifically, the number of trees in the forest (n_estimators), the maximum number of features to consider when searching for the best split (max_features), and the maximum depth of the trees (max_depth) were tuned. By careful tuning of these hyperparameters, we negate the need for traditional feature selection, as this is then implicit in the Random Forest algorithm when hyperparameters are set to appropriate values. The random seed was set to a random integer generated by the Numpy.random library [62]. Other hyperparameters relating to limiting the size of the trees (other than maximum depth) were left at default values as resources were sufficient to compute unpruned trees in reasonable time.

We validated our models using a student-level k-fold cross-validation (k=2). For each iteration of the classifier, a random 50% of students were assigned to the training set, the remaining 50% to the test set, the process was repeated with the sets reversed, and results computed after pooling predictions across the folds. By using a low k value, we increase the size of our test set, increasing the likelihood that successful models will generalize to new data. This process was repeated for 15 iterations and the results were averaged across iterations. We selected 15 iterations to balance computation time and reliability across multiple training/testing pairs. Although setting k=2 imposes a stringent test of the model by removing half the data for the test set, it helps to ensure that the models will generalize to new students.

We note that for some of the latent classes there is a substantial data skew (more instances of not graduating than graduating). Class imbalance poses a challenge because supervised learning methods tend to bias predictions towards the majority class. To compensate for this concern, we used the SMOTE algorithm [12] to create synthetic instances of the minority class by interpolating feature values between an instance and its randomly chosen nearest neighbors until the classes were equated. SMOTE was only applied on the training sets; the original class distributions were maintained in the test sets in order to ensure validity of the results.

5. RESULTS

We report area under the receiver operating characteristic curve (AUROC). Whereas overall recognition rate/model accuracy is susceptible to data skew, AUROC presents the result relative to chance (0.5).

5.1 Generalization

We first compared how models trained on each class individually compared to a model trained upon all data. The all data model was evaluated using a slicing analysis, where predictive model performance is evaluated by "slicing" along subpopulations, in this case, the computed LCA classes. We trained a random forest classifier on all of the data (41,359 instances) and then evaluated it by the five latent classes previously identified. The results of this analysis are shown in Figure 1 with the baseline reflecting chance performance (AUROC of 0.5).

Each of the individual models performed above chance, suggesting that our methods generalized across classes. However, there was a disparity between the classes, the highest performing (class 3, white high SES) performed 10% better than the lowest performing (class 4, Asian high SES). The two lowest AUROC scores were for class 1 (Black, Mid SES) and class 4 (Asian, high SES); the two classes with the lowest number of instances (1,745 and 1,051 respectively).

Our model trained on all students performed better across all classes than individually trained models, with improvements ranging from 6-10%. The difference between the worst performing and best performing groups also decreased to 6%, implying better generalization across groups. One reason for this may be the improved power that comes with a higher number of instances; the all data model was trained on 41,369, instances, more than double that of the largest LCA class (16,959 instances).

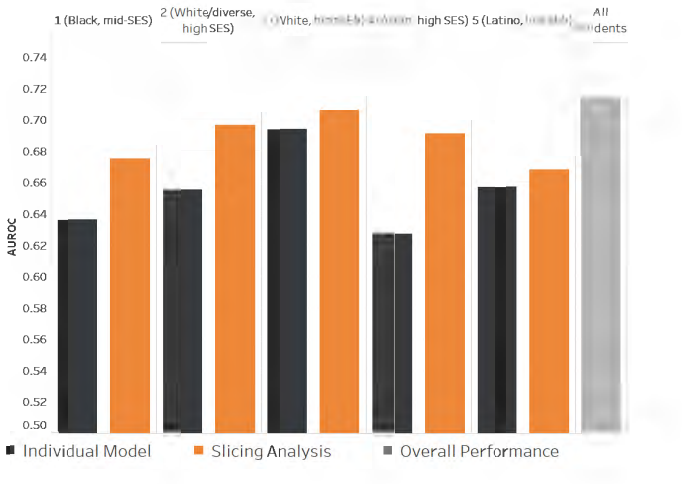

Figure 1. Accuracy of individual models compared to a slicing analysis of a classifier trained on all data

5.2 Fairness

We next examined the fairness of our model. Using the model trained on all data (41,359 instances) we computed five ROC curves, one for each of the latent classes. Recall that this was done for 15 iterations with k=2 cross validation. The ROC curves for a single iteration is shown in Figure 2, this iteration was chosen as the ABROCA scores are similar to the averages shown in Table 3. A sixth ROC curve for all students is also shown for comparison.

In order to formally compare two curves, we use Absolute Between ROC Area (ABROCA) [26], defined as:

$$\int_0^1 |ROC_b(t) - ROC_c(t)|\, dt$$

Here, ROC_b is the baseline curve and ROC_c is the comparison curve. ROC curves characterize model accuracy as the likelihood of correct positive predictions versus the likelihood of false positive predictions. ABROCA measures the absolute difference between two curves, allowing for the possibility that the curves may cross each other (see [26] for details). A higher ABROCA value between two groups implies a higher difference in predictions and thus more unfairness in the model.

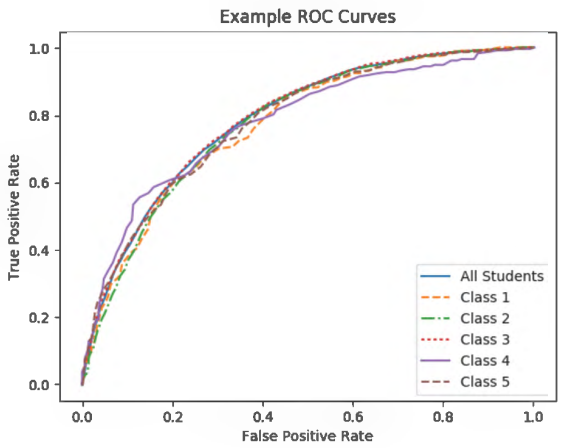

Figure 2. Sample ROC curves for each Latent class from model trained on 41,359 instances

We used class 3 (White, high SES) as the baseline ROC curve as it is the highest performing group in our dataset. It is also a group typically overrepresented in Educational Data Mining [7]. When compared to this class, the other classes had low ABROCA values (see Table 3), perhaps unsurprising given the similarity of the curves in Figure 2. In general, the ABROCA values were all low with only small differences between classes, leading us to conclude that our model was providing fair predictions across our sociodemographic groups.

Table 3. Slicing analysis by latent class from model trained on all data

LCA Class	AUROC	ABROCA
1 (Black, mid-SES)	0.675	0.011
2 (White/diverse, high SES)	0.696	0.005
3 (White, high SES)	0.706	-
4 (Asian, high SES)	0.691	0.016
5 (Latino, low SES)	0.668	0.008

6. DISCUSSION

On-time 4-year college graduation is something of a "holy-grail" for students, parents, and educators alike [58]. Although efforts to improve college enrollment have been paying off, graduation rates are still lackluster with troublesome achievement gaps stubbornly persisting. Big data approaches might offer a potential solution to improving college graduation rates by providing new insights into the "ingredients" of success. However, they have their own set of limitations and biases, which need to be addressed before we uncover their full potential. Accordingly, we investigated how

predictive models of 4-year college graduation generalized across sociodemographic subgroups identified through latent class analysis and whether they yielded fair predictions for the different groups of students.

6.1 Main Findings

Whilst much of the previous work has relied on limited datasets and traditional statistical techniques [53, 61, 69], we harness a large and diverse dataset with greater potential for generalizability. Specifically, using data from students' college application, we have been able to predict college graduation with moderate accuracy across demographic subgroups. We also show through slicing analysis that a model trained on all data generalizes across all of the subgroups and outperforms individual models (improvement ranging from 6-10%). Training a model on all of the data also reduced the disparity between the subgroups.

By evaluating the ABROCA metric, we were also able to examine which of our subgroups (if any) the classifier was treating unfairly. An unfair model would perform generate less accurate predictions for a given subgroup compared to the baseline group (White High SES in our case). In general, the differences in ABROCA scores were small, suggesting that our model treats no one class significantly different from another.

Whilst all of our models' predictions were substantially more accurate than a chance, there were still inaccuracies. In many ways, this result is reassuring, as we have only considered data from high school. The error that exists across all of our models confirm that college success does not merely depend on a student's environment, past achievement, and experiences. What students experience and do in college plays a critical role in their success. Simply put, there is no predetermination. This gives us hope that through careful data mining we can soon begin to close the achievement gaps that exist across different sociodemographics.

6.2 Applications

It is perhaps easiest to start with how these models should *not* be used. Specifically, the models should not be used to make college admissions decisions because their accuracy scores are insufficient to drive life-changing decisions for individual students and they do not capture several additional factors of the college years that are important for success (e.g. financial needs, life-altering events, social pressures).

We show that it is possible to build generalizing and fair detectors in this domain. On a larger scale, we hope to use this research to provide actionable advice for educators so that they may better prepare students for college success. Further analysis is needed to derive these personalized recommendations, especially since the current models are correlational and thereby unsuited for causal inference.

There are further applications at the college level. Many U.S. colleges have committed to improving 4-year graduation rates [13]. This has resulted in an increased reliance on educational data mining approaches, especially methods to identify "at-risk" students early on [2]. A common issue however is that early warnings are not early enough [5]. Our models consider college application data, so enable us to pre-identify students who might need additional support before they begin their studies. Of course, the models' assessments should be privately communicated to the student's themselves and perhaps to a trusted counselor so they are empowered to take whatever next step is in their best interests.

6.3 Limitations and Future Work

All studies have limitations and ours is no exception. Each of the latent classes had different graduation rates and varied number of instances (a difference of 15,909 instances between the smallest and the largest groups). We attempted to account for class imbalance via synthetic oversampling. However, further work is required to evaluate how the number of instances influenced our results. Future work will also explore the effect of increasing the amount of data used to train models.

Second, our sample only included students who applied to schools that accepted the Common App, which would introduce selection bias, which we cannot account for in this work. Further study is required to investigate how the results generalize to other colleges in the U.S. and beyond.

In addition to addressing these limitations, there are also several promising avenues for future work. For example, since biased variables seem to be predictive in this domain, we will also look into ways to create fair models without fully ignoring the biased variables, perhaps by deriving unbiased proxies.

Our models utilized a range of features including socioeconomic factors, academic history, cognitive ability, the high school environment, and indicators of extracurricular participation that may reflect non-cognitive characteristics. Previous work using the CommonApp-NSC dataset work has shown that different feature groups [33] achieve different classification accuracies. In future work we intend to explore fairness for different feature groups and incorporate insights into the design of fairer models.

When computing the LCA classes, we did not include gender as a variable. However, gender might be more relevant when it comes to specialized outcomes, such as STEM graduation where there are significant disparities across the genders. Relatedly, we also will explore other outcome metrics such as 6-year graduation and STEM graduation, an area with wide achievement gaps when it comes underrepresented groups [56].

6.4 Concluding Remarks

In conclusion, the age of big data brings with it big opportunity, and big responsibility. Although a predictive modeling approach applied to big data has considerable potential in providing new insights to illuminate persistent challenges, these methods have own weaknesses, particularly when it comes to making biased predictions. Thus, we must also consider how our models are perpetuating pre-existing bias and how this can be prevented. Taking the case of predicting on-time college graduation outcomes, we show that our models both generalize and are fair to various sociodemographics subgroups, a critical step towards using these models more broadly.

ACKNOWLEDGMENTS

We are grateful to the Common App for providing us with the dataset, which made this work possible. We also thank Parker Goyer for her assistance in coding the NSC data and to Tammer Ibrahim for his help with the computational infrastructure. This research was supported by the Walton Family Foundation, the Mindset Scholars Network, the Bill & Melinda Gates Foundation, the Joyce Foundation, the Overdeck Family Foundation, and the Raikes Foundation. Any opinions, findings and conclusions, or recommendations expressed in this paper are those of the authors and do not necessarily reflect the views of the funding agencies.

REFERENCES

[1] Allen, D. 1999. Desire to finish college : an empirical link between motivation and persistence. *Research in Higher Education*. 40, 4 (1999), 461–485. DOI:https://doi.org/10.1023/A:1018740226006.

[2] Allensworth, E. and Easton, J.Q. 2007. What matters for staying on-track and graduating in chicago public high schools: a close look at course grades, failures, and attendance in the freshman year. *Consortium on Chicago School Research*. (2007), 1–61.

[3] Astin, A.W. 1977. *Four critical years. effects of college on beliefs, attitudes, and knowledge.* Jossey-Bass Publishers.

[4] Baker, R.S.J. d. and Gowda, S.M. 2010. An analysis of the differences in the frequency of students' disengagement in urban, rural, and suburban high schools. *Proceedings of the 3rd International Conference on Educational Data Mining*. (2010), 11–20.

[5] Beaudoin, B. and Kumar, P. 2012. Using data to identify at-risk students and develop retention strategies. *EAB Custom Research Brief*. (2012).

[6] Big Risks, Big Opportunities: the Intersection of Big Data and Civil Rights | whitehouse.gov: 2016. *https://obamawhitehouse.archives.gov/blog/2016/05/04/big-risks-big-opportunities-intersection-big-data-and-civil-rights*. Accessed: 2019-02-22.

[7] Blanchard, E.G. 2012. On the weird nature of its/aied conferences: a 10 year longitudinal study analyzing potential cultural biases. *Intelligent Tutoring Systems* (2012), 280–285. DOI:https://doi.org/10.1007/978-3-642-30950-2_36.

[8] Bosch, N. et al. 2016. Using video to automatically detect learner affect in computer-enabled classrooms. *ACM Transactions on Interactive Intelligent Systems* (2016), 1–26. DOI:https://doi.org/10.1145/2946837.

[9] Bowers, E.P. et al. 2010. The five cs model of positive youth development: a longitudinal analysis of confirmatory factor structure and measurement invariance. *Journal of Youth and Adolescence*. 39, 7 (2010), 720–735. DOI:https://doi.org/10.1007/s10964-010-9530-9.

[10] Breiman, L. 2001. Random forests. *Machine Learning*. 45, 1 (2001), 5–32. DOI:https://doi.org/10.1023/A:1010933404324.

[11] Carnevale, A.P. et al. 2013. *The college payoff: education, occupations, lifetime earnings*. Center on Education and the Workforce.

[12] Chawla, N. V. et al. 2002. SMOTE: synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*. 16, 1 (Jun. 2002), 321–357. DOI:https://doi.org/10.1613/jair.953.

[13] Complete College America 2014. *Four-year myth*. Complete College America.

[14] Crawford, K. and Calo, R. 2016. There is a blind spot in ai research. *Nature*. (2016). DOI:https://doi.org/10.1038/538311a.

[15] Dastin, J. 2018. Amazon scraps secret ai recruiting tool that showed bias against women. *Reuters*.

[16] Duckworth, A.L. and Allred, K.M. 2012. Temperament in the classroom. *Handbook of temperament*. Guilford Press New York, NY. 627–644.

[17] Duckworth, A.L. and Yeager, D.S. 2015. Measurement matters: assessing personal qualities other than cognitive ability for educational purposes. *Educational Researcher*. 44, 4 (2015), 237–251. DOI:https://doi.org/10.3102/0013189X15584327.

[18] Dutt, A. et al. 2017. A systematic review on educational data mining. *IEEE Access*. 5, (2017), 15991–16005. DOI:https://doi.org/10.1109/ACCESS.2017.2654247.

[19] Dweck, C.S. 2006. *Mindset: the new psychology of success*. Random House.

[20] Dwork, C. et al. 2012. Fairness through awareness. *Proceedings of the 3rd innovations in theoretical computer science conference* (2012), 214–226. DOI:https://doi.org/10.1145/2090236.2090255.

[21] Eccles, J.S. et al. 2003. Extracurricular activities and adolescent development. *Journal of Social Issues*. 59, 4 (2003), 865–889. DOI:https://doi.org/10.1086/223736.

[22] Eckland, B.K. 1964. Social class and college graduation : some misconceptions corrected. *American Journal of Sociology*. 70, 1 (1964), 36–50. DOI:https://doi.org/10.1086/223736.

[23] Epstein, J.L. and Hollifield, J.H. 2005. Title i and school--family--community partnerships: using research to realize the potential. *Journal of Education for Students Placed at Risk (JESPAR)*. (2005). DOI:https://doi.org/10.1207/s15327671espr0103_6.

[24] Ewert, S. 2012. Fewer diplomas for men: the influence of college experiences on the gender gap in college graduation. *The Journal of Higher Education*. 83, 6 (2012), 824–850. DOI:https://doi.org/10.1353/jhe.2012.0042.

[25] Gajane, P. and Pechenizkiy, M. 2017. On formalizing fairness in prediction with machine learning. *arXiv preprint arXiv:1710.03184*. (2017).

[26] Gardner, J. et al. 2019. Evaluating the fairness of predictive student models through slicing analysis. *Proceedings of the 10th Conference on Learning Analytics and Knowledge* (Tempe, AZ, USA, 2019), 10. DOI:https://doi.org/10.1145/3303772.3303791.

[27] Gardner, M. et al. 2008. Adolescents' participation in organized activities and developmental success 2 and 8 years after high school: do sponsorship, duration, and intensity matter? *Developmental psychology*. 44, 3 (2008), 814–830. DOI:https://doi.org/10.1037/0012-1649.44.3.814.

[28] Goodwin, B. and Hein, H. 2016. Research says/the x factor in college success. *Educational Leadership*. 73, 6 (2016), 77–78.

[29] High school demographics continue to impact college success: 2016. *https://studentclearinghouse.org/blog/high-school-demographics-continue-to-impact-college-success/*.

[30] Ho, T.K. 1995. Random decision forests. *Proceedings of the Third International Conference on Document Analysis and Recognition* (Washington, DC, USA, 1995), 278–282.

DOI:https://doi.org/10.1109/ICDAR.1995.598994.

[31] Holland, A. and Andre, T. 1987. Participation in extracurricular activities in secondary school: what is known, what needs to be known? *Review of Educational Research*. 57, 4 (1987), 437–466. DOI:https://doi.org/10.3102/00346543057004437.

[32] Humbert, M.L. et al. 2006. Factors that influence physical activity participation among high-and low-ses youth. *Qualitative health research*. 16, 4 (2006), 467–483. DOI:https://doi.org/10.1177/1049732305286051.

[33] Hutt, S. et al. 2018. Prospectively predicting 4-year college graduation from student applications. *Proceedings of the 8th International Conference on Learning Analytics and Knowledge* (New York, NY, USA, NY, USA, 2018), 280–289. DOI:https://doi.org/10.1145/3170358.3170395.

[34] Ishitani, T.T. 2006. Studying attrition and degree completion behavior among first-generation college students in the united states. *The Journal of Higher Education*. 77, 5 (2006), 861–885. DOI:https://doi.org/10.1353/jhe.2006.0042.

[35] Kanno, Y. and Cromley, J.G. 2013. English language learners' access to and attainment in postsecondary education. *TESOL Quarterly*. (2013). DOI:https://doi.org/10.1002/tesq.49.

[36] Kusner, M.J. et al. 2017. Counterfactual fairness. *Advances in Neural Information Processing Systems 30* (2017), 4066–4076.

[37] Larson, R.W. et al. 2006. Differing profiles of developmental experiences across types of organized youth activities. *Developmental psychology*. 42, 5 (2006), 849–863. DOI:https://doi.org/10.1037/0012-1649.42.5.849.

[38] Marsh, H. and Kleitman, S. 2002. Extracurricular school activities: the good, the bad, and the nonlinear. *Harvard Educational Review*. 72, 4 (2002), 464–515. DOI:https://doi.org/10.17763/haer.72.4.051388703v7v7736.

[39] McFarland, J. et al. 2017. The condition of education 2017. *National Center for Education Statistics*. (2017).

[40] Muthén, L.K. and Muthén, B.O. 2015. *Mplus. seventh edition.*

[41] NCES 2016. Digest of education statistics 2016. table 326.10. graduation rate from first institution attended for first-time, full-time bachelor's degree- seeking students at 4-year postsecondary institutions. *Digest Of Education Statistics*. U.S. Department of Education, National Center for Education Statistics.

[42] Ocumpaugh, J. et al. 2014. Population validity for educational data mining models: a case study in affect detection. *British Journal of Educational Technology*. (2014). DOI:https://doi.org/10.1111/bjet.12156.

[43] Pedregosa, F. et al. 2011. Scikit-learn: machine learning in python. *Journal of Machine Learning Research*. 12, (2011), 2825–2830.

[44] Ver Ploeg, M. 2002. Children from disrupted families as adults: family structure, college attendance and college completion. *Economics of Education Review*. (2002). DOI:https://doi.org/10.1016/S0272-7757(00)00050-9.

[45] Russell, C. et al. 2017. When worlds collide: integrating different counterfactual assumptions in fairness. *Advances in Neural Information Processing Systems 30* (2017), 6414--6423.

[46] Ryan S.J.d. Baker, K.Y. 2009. The state of educational data mining in 2009: a review and future visions. *Journal of Educational Data Mining*. (2009). DOI:https://doi.org/10.1109/ASE.2003.1240314.

[47] Sable, J. and Plotts, C. 2010. Documentation to the nces common core of data public elementary/ secondary school universe survey: school year 2007–08 (nces 2010-302rev). (2010).

[48] Saegert, S.C. et al. 2007. *Report of the APA Task Force on Socioeconomic Status APA Task Force on Socioeconomic Status*.

[49] Samei, B. et al. 2015. Modeling classroom discourse: do models that predict dialogic instruction properties generalize across populations? *Educational Data Mining* (2015), 444–447.

[50] Sculley, D. et al. 2018. Winner's curse? on pace, progress, and empirical rigor. *International Conference on Learning Representations* (2018).

[51] Sigle-Rushton, W. and McLanahan, S. 2004. Father absence and child well-being: a critical review. *The future of the family*. (2004).

[52] Snyder, T.D. et al. 2016. *Digest of education statistics, 2015*. National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education.

[53] Stumpf, H. and Stanley, J.C. 2002. Group data on high school grade point averages and scores on academic aptitude tests as predictors of institutional graduation rates. *Educational and Psychological Measurement*. 62, 6 (2002), 1042–1052. DOI:https://doi.org/10.1177/0013164402238091.

[54] Tate, W.F. 2006. Race-ethnicity, ses, gender, and language proficiency trends in mathematics achievement: an update. *Journal for Research in Mathematics Education*. (2006). DOI:https://doi.org/10.2307/749636.

[55] The Common Application: *http://www.commonapp.org/*. Accessed: 2017-09-27.

[56] Thompson, R. and Bolin, G. 2011. Indicators of success in stem majors: a cohort study. *Journal of College Admission*. (2011).

[57] Tinto, V. 2008. Moving beyond access: college success for low-income, first-generation students. *The Pell Institute for the Study of Opportunity in Higher Education*. (2008).

[58] US Department of Education, N. 2018. The condition of education - 2018. *Institute of Education Sciences*. (2018).

[59] Vidyasagar, M. 2002. *A theory of learning and generalization*. Springer-Verlag.

[60] Walpole, M. 2008. Emerging from the pipeline : african american students , socioeconomic status , and college experiences and outcomes. *Research in Higher Education*. 49, 3 (2008), 237–255. DOI:https://doi.org/10.1007/slll62-007-9079-y.

[61] Walpole, M. 2003. Socioeconomic status and college: how ses affects college experiences and outcomes. *The Review*

of Higher Education. 27, 1 (2003), 45–73. DOI:https://doi.org/10.1353/rhe.2003.0044.

[62] van der Walt, S. et al. 2011. The numpy array: a structure for efficient numerical computation. *Computing in Science & Engineering*. 13, 2 (Mar. 2011), 22–30. DOI:https://doi.org/10.1109/MCSE.2011.37.

[63] Walton, G.M. and Cohen, G.L. 2011. A brief social-belonging intervention improves academic and health outcomes of minority students. *Science (New York, N.Y.)*. 331, 6023 (Mar. 2011), 1447–1451. DOI:https://doi.org/10.1126/science.1198364.

[64] Warren, J.R. et al. 2011. Employment during high school: consequences for students' grades in academic courses. *Journal of Vocational Education Research*. (2011). DOI:https://doi.org/10.5328/jver26.3.366.

[65] Warren, J.R. 2002. Reconsidering the relationship between student employment and academic outcomes: a new theory and better data. *Youth & Society*. 33, 3 (2002), 366–393. DOI:https://doi.org/10.1177/0044118X02033003002.

[66] Waugh, G. et al. 1994. Using ethnicity, sat/act scores, and high school gpa to predict retention and graduation rates. *Florida Association for Institutional Research Conference*. (1994).

[67] Zou, J. and Schiebinger, L. 2018. AI can be sexist and racist — it's time to make it fair. *Nature*. (2018). DOI:https://doi.org/10.1038/d41586-018-05707-8.

[68] Zwick, R. 2002. Is the sat a "wealth test"? *Phi Delta Kappan*. 84, 4 (2002), 307–311. DOI:https://doi.org/10.1177/003172170208400411.

[69] Zwick, R. and Sklar, J.C. 2005. Predicting college grades and degree completion using high school grades and sat scores: the role of student ethnicity and first language of student ethnicity and first language. *American Educational Research Journal*. 42, 3 (2005), 439–464. DOI:https://doi.org/10.3102/00028312042003439.