Using Machine Learning and Genetic Algorithms to Optimize Scholarship Allocation for Student Yield

Lovenoor Aulck
University of Washington
Seattle, WA, USA
laulck@uw.edu

Dev Nambi
F.H. Cancer Research Center
Seattle, WA, USA
dnambi@fredhutch.org

Jevin West
University of Washington
Seattle, WA, USA
jevinw@uw.edu

ABSTRACT

Effectively estimating student enrollment and recruiting students is critical to the success of any university. However, despite having an abundance of data and researchers at the forefront of data science, universities are not fully leveraging machine learning and data mining approaches to improve their enrollment management strategies. In this project, we use data at a large, public university to increase their student enrollment. We do this by first predicting the enrollment of admitted first-year, first-time students using a suite of machine learning classifiers (AUROC = 0.85). We then use the results from these machine learning experiments in conjunction with genetic algorithms to optimize scholarship disbursement. We show the effectiveness of this approach using actual enrollment metrics. Our optimized model was expected to increase enrollment yield by 15.8% over previous disbursement strategies. After deploying the model and confirming student enrollment decisions, the university actually saw a 23.3% increase in enrollment yield. This resulted in millions of dollars in additional annual tuition revenue and a commitment by the university to employ the method in subsequent enrollment cycles. We see this as a successful case study of how educational institutions can more effectively leverage their data.

KEYWORDS

education, funding, tuition, university, college, machine learning, genetic algorithm

ACM Reference Format:
Lovenoor Aulck, Dev Nambi, and Jevin West. 2019. Using Machine Learning and Genetic Algorithms to Optimize Scholarship Allocation for Student Yield. In *SIGKDD '19: ACM SIGKDD Conference on Knowledge Discovery and Data Mining, August 4–8, 2019, Anchorage, AK*. ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/1122445.1122456

1 INTRODUCTION

Managing student enrollment is one of the core administrative tasks of any university. However, it is far from simple as universities aim to attract and retain the best students with limited resources [4, 13]. Enrollment management has wide-ranging implications on institutions' student body composition as well as their budgeting and finances, where a reliance on tuition income necessitates accurately forecasting student enrollments [12, 27]. One instrument that has continually been leveraged in the pursuit of enrollments and the associated tuition income is financial aid as receiving a financial aid award increases the likelihood of a student enrolling [10, 13, 16]. While financial aid remains a powerful mechanism for institutions to reach their admissions and revenue targets, miscalculating projected student enrollments and mismanaging financial aid funds can have severe implications (such as rescinding over-committed offers[1])[2]. Furthermore, as institutions face tightening budgets and find their pricing policies continually under scrutiny, it remains imperative for them to optimize the resources they have by maximizing enrollments and the associated tuition revenue from financial aid programs [6, 7, 11, 15]. As such, accurately predicting enrollment and optimizing how student aid is disbursed is critical to enrollment management with financial implications that cascade across the entirety of an institution. In this work, we develop an approach to address this challenge, implemented it for a recent entering class, and found that it far outperformed previous strategies.

Predicting enrollment and optimizing the allocation of student aid requires data on student admissions, operational expenses, and budgets. This data is stored in institutions' organizational databases or can be extracted from historical and operational records. However, despite having this abundance of data on previous enrollments and finances, institutions are often slow to leverage it to gain actionable insights and improve institutional processes [17, 23, 30]. What's more, using data for insights in education is less prevalent at traditional campuses (i.e. schools where learning is primarily on-campus) and more common in online and computerized environments, which are much more amenable to the collection and analysis of digitized data [20]. To this end, traditional universities remain "data-rich" but are "information-poor" in that they have the raw data needed to extract intelligible insights but are unable to do so due to infrastructure limitations and untrained personnel, among other reasons [25]. This results in the outsourcing of data-centric enrollment work (including enrollment prediction and developing scholarship disbursement strategies) to full-service consulting firms, which do not disclose their proprietary approaches or how their results are evaluated [14]. The lack of motivation for consulting services to disseminate their work coupled with institutions trying to maintain competitive advantages in recruitment limits the extent of published research on how institutions can more effectively utilize data in enrollment management to improve existing processes. As a result, this dearth of literature provides little to demonstrate how data mining and machine learning can assist in the critical mission of enrollment management and in allocating financial aid.

[1] See https://bit.ly/2Scxqj6 as a recent example.

SIGKDD '19, August 4–8, 2019, Anchorage, AK

ACM ISBN 978-1-4503-9999-9/18/06…$15.00
https://doi.org/10.1145/1122445.1122456

In this project, we mine data from a large, public university in the United States (US) to optimize the disbursement of a merit-based scholarship for domestic non-resident students. We do this in two steps. We create a predictive model of student enrollment. We then use a genetic algorithm to optimize scholarship disbursement to maximize student enrollment based on this predictive enrollment model. We conducted this work during the most recent admissions cycle of the university and the optimized awards were given to the latest entering class. After seeing improvement in student enrollment yield and an increase of millions of dollars in annual tuition revenue, the university incorporated our approach into their enrollment management process. We believe this project is a case study for other institutions seeking to similarly leverage institutional data for improving enrollment forecasting and financial aid allocation.

2 RELEVANT WORK

The following discussion of relevant work is not exhaustive with respect to enrollment prediction and financial aid optimization. It is intended to give examples of relevant approaches with a focus on more recent work. While there is some work showing how to predict enrollment, there is very little showing how to allocate scholarships and hardly anything that ties the two together.

2.1 Predicting Enrollment

A few studies have employed machine learning and data mining techniques to predict enrollment at a university using non-neural approaches. DesJardins developed a logistic regression model using a dataset of approximately 14,400 students from an undisclosed tier I research university in the US Midwest. DesJardins' model gave an area under the receiver operating characteristic curve (AUROC) of 0.72 when predicting whether or not a student will enroll [5]. Similarly, Goenner and Paul used logistic regression to predict which of over 15,000 students at a medium-sized US university would eventually enroll [9]. With a highly imbalanced dataset, their regression model gave an AUROC value of 0.87. Nandeshwar and Chaudhari later used a suite of learners, including Naive Bayes and tree-based models, to predict which of approximately 28,000 students would enroll at West Virginia University [19]. They were interested in variables contributing to students' decisions (finding financial aid to be an important factor) and did not give an assessment of how well their models fared outside of accuracy (which was about 84%).

In addition to the above studies examining non-neural approaches for predicting enrollment, some studies have also found that neural approaches fare very well for the same task and often perform better than non-neural approaches. For example, Walczak evaluated different neural network designs when examining predictions of student enrollment at a small US private liberal arts college, stressing the problem as one of resource allocation [28]. Using a few thousand students, Walczak found that backpropagating neural networks fared best among those compared. Walczak and Sicich later compared neural networks versus logistic regression to predict whether students would enroll at a given institution at both a small US private university and at a large public US university [29], finding that neural networks performed better than logistic regression. Chang used logistic regression, decision trees, and neural networks to predict the enrollment of admitted applicants at an undisclosed university, also finding that neural networks outperformed the other models when judging by classification accuracy [3]. Recently, Shrestha et al. looked to predict whether undergraduate and graduate international students admitted to an undisclosed Australian university would enroll [24]. Their approach included looking at Naive Bayes, decision trees, support vector machines, random forests, K-nearest neighbors, and neural networks. In their setup, logistic regression and neural networks fared best. It should be noted that there is a scarcity of literature among the works listed above on using ensemble approaches in predicting student enrollment and comparing their performance to neural approaches.

2.2 Scholarship Optimization

While there are some examples of works examining the use of machine learning in predicting enrollment, there is very little detailing scholarship disbursement strategies, especially ones leveraging machine learning and/or numerical optimization techniques. One example is the work of Alhassan and Lawal, who demonstrated the use of tree-based models for determining which students would be awarded scholarships in Nigeria [1]. Alhassan and Lawal describe the results as "effective" and "efficient" compared to approaches previously used but did not provide more on the success of the disbursement strategy. Spaulding and Olswang demonstrated the use of discriminant analysis to model the enrollment decisions of students based on varying need-based financial aid awards at an undisclosed university in the US [26]. They found that changes in their award policy would yield only small upticks in enrollment.

One work used machine learning to predict enrollment in conjunction with a numerical optimization technique to disburse scholarships. Sarafraz et al. used neural networks with genetic algorithms to optimize financial aid allocations and while our research is similar in spirit, there are a few notable differences [22]. Firstly, the scholarship fund optimized in this work is merit-based, meaning there are upper and lower bounds on scholarship awards that are specific to each student. This makes for a more difficult optimization task. We also examine alternative predictive models beyond just neural networks (such as ensemble approaches) and use a larger dataset in terms of both the number of observations and the number of features (over 72,000 observations vs 4,082; over 100 features vs 6). We also provide a comprehensive description of final model performance across multiple metrics and a detailed outline of how genetic algorithms can be used for aid disbursement, including a binning framework to drive the optimization task. Finally, we share real-world enrollment metrics after employing the scholarship optimization to demonstrate the effectiveness of our approach.

3 METHODS

We present the methods for this work in the following order: first, we give an overview of the setting for this research; then, we describe the data as well as feature engineering performed on the data; we then describe the process for predicting enrollment; finally, we discuss optimization constraints and outline the process for scholarship optimization.

3.1 Setting

This scholarship optimization work was performed at a large, public US University (the University[2]) in early 2018. The scholarship fund examined was created to maintain the University's academic standards while maximizing the enrollment of first-time, first-year (freshmen) domestic non-resident (DNR) students by giving them financial incentive to enroll at the University. DNR students are students from the US who are not from the state in which the University is located. DNR students account for significantly larger tuition charges than their resident (i.e. in-state) counterparts and, therefore, their enrollment is of high importance from a budgeting and finance perspective. Tens of millions of dollars in total are awarded annually to these students as part of the scholarship fund with millions eventually spent each year on students who enroll.

The scholarship fund that we examined (DNR scholarships) was designated to be disbursed in a merit-based manner. As such, students with higher academic profiles, as defined later, were given equal or larger scholarships than those with lower academic profiles, regardless of financial need. Additionally, only freshmen DNR students who were accepted to the University were eligible to receive a DNR scholarship award. All admitted DNR students were automatically considered for a DNR scholarship and students did not need to apply for the scholarship separately.

In years prior, developing the disbursement strategy for the DNR scholarship was outsourced to external consulting services. For the last full application cycle (the 2018 entering class), it was brought under the technical stewardship of the University. This is the application cycle for which we optimized scholarship disbursement. It should be noted that the models that were previously developed for the disbursement of this scholarship fund were proprietary to the consulting service and could not be leveraged in any way. However, student application, enrollment, and scholarship data from prior years was available. When describing results, we compare the results from our approach to that developed by the consulting service. We cannot compare the approach detailed in this writing to a completely unoptimized approach or one that is randomized.

Award-receiving students concurrently learned of the amount of their scholarship and of their admittance to the University. However, not all applications were scored by admissions officers when the first round of awards were to be given. This was primarily due to the time taken to review tens of thousands of admissions applications and typical review timelines at the University. We did not know of every admitted student at the time of optimization yet the scholarship awards were only to be given to admitted students. Thus, the last full application cycle's data could not be used directly in the optimizations. Instead, our approach used data from previous years to develop a fund allocation strategy and then apply this strategy to the last application cycle. This was with the expectation that applicants in the last application cycle were similar to years prior and we checked to ensure that this was the case.

3.2 Data

The data for this work consisted of information on all freshmen DNR applicants to the University from 2014-2017 with usable data. This totaled 72,589 students. Data from the study came from two major sources, both of which were regularly maintained by the University: the students' admissions applications and their Free Application for Federal Student Aid (FAFSA) information. The FAFSA is an application prepared by incoming and current US college students to determine their eligibility for financial aid. It should be noted that no additional data was collected for this project. Examples of data pulled from students' admissions applications included their high school courses taken, entrance exam scores, college GPA (if they had taken classes for credit), whether they received an athletic scholarship, whether they were a first-generation college student, and their parents' educational attainment. These were all self-reported and verified by the University as needed. Data directly from and derived from student FAFSA filings included students' family income, their expected family contribution to college expenses (as calculated by the University), and institutional loan amounts awarded to the student. Also included in the data were indicators of whether each student was accepted to the University and whether the student eventually enrolled. Of the 72,589 students in the dataset, 34,874 were admitted (48.04% of all) and 5,081 enrolled (14.57% of admitted, 7.00% of all). Demographic variables such as gender, race, and ethnicity were available but were not included in the data as discussed in Section 4.1.

Within the data were values on tuition amounts students would pay on an annual basis, their financial aid grants and scholarships awarded (outside of DNR scholarship awards), and their DNR scholarship award amount. These variables were not included in any prediction or optimization model on their own. Instead, we created a "`reduced_tuition`" variable which was the annual tuition amount for the students less their total grants and scholarships (i.e. the other two variables summed). We used this variable as a single financial aid and tuition-related feature for the predictions and optimizations discussed below.

DNR applicants to the University were on average 18.0 years old at the time of application. About 17% of applicants had taken part in a college in high school program but about 99.5% of applicants were applying as freshmen entrants, meaning they were below the credit threshold to be considered sophomores upon entry to the University. About 66% of applicants had filled a FAFSA.

3.3 Feature Engineering

Prior to prediction and optimization, we engineered features from existing variables. First, we either converted categorical variables to dummy variables or replaced them with a binary indicator variable. Then, we grouped students based on their FAFSA award amounts into 6 discrete bins, each of which was used as a categorical feature. We created binary indications of whether students attended each of the 10 most popular high schools for student applications and did the same for the 10 most popular states from which students applied. A binary indication was also created for a student athlete designation as each sport had its own application codes. In addition, we also created a separate binary indication for whether the student was transferring any credits from a college in high school program. Students' academic interests were also pulled from their applications and were grouped into 12 broader categories. We then created binary indications of whether a student was interested in each of the categories. Only students' first application to the University

[2]University administrative offices requested that the institution not be identified.

and the resulting admissions/enrollment decisions were included in the data. This resulted in a total of 108 features. Not all applicants filed a FAFSA form and we imputed missing FAFSA-related values using gradient boosted regression trees [8].

3.4 Predicting Enrollment

To predict enrollment, we first randomly divided the data using a 80-20 training-test split, with 57,359 students in the training set and 14,340 students in the test set. We did not re-balance the data with respect to classes. We scaled the training data by subtracting the median of each feature and dividing by the feature's interquartile range. We subsequently scaled the test data using the scaling values from the training data. The binary outcome variable indicating whether the student enrolled at the University was not scaled.

After performing the training-test split, we trained 7 machine learning (ML) classifiers on the training set to predict enrollment. These classifiers were: a bagging tree ensemble (BC), gradient boosted trees (XGB), K-nearest neighbors (KNN), random forests (RF), regularized logistic regression (LR), support vector machines (SVM), and a neural network with 3 hidden layers (MLP). We tuned the hyperparameters for each of the classifiers using 5-fold cross validation on the training set. We report performance from all classifiers on the test set, which was not used to train the classifiers and only used to evaluate final performance. We used the classifier with the best performance to optimize aid disbursement.

3.5 Modeling Constraints

Several constraints were posed on the scholarship disbursement in accordance with the strategic goals University administrators. These constraints underwent many changes during the modeling process, not all of which will be discussed. Due to University policy, exact values for awards and budgets will also not be discussed. That said, the constraints on the disbursement strategy were as follow, where F represents funds in DNR scholarship offers, B represents funds in the DNR scholarship budget, N specifies a count of students, and S specifies a scholarship award amount:

(1) The total amount spent on DNR scholarships (F_{spent}) cannot exceed a pre-determined amount (B_{spent}): $F_{\text{spent}} \leq B_{\text{spent}}$
(2) The total amount offered to students in DNR scholarships regardless of whether they enroll (F_{offered}) cannot exceed a pre-determined amount (B_{offered}): $F_{\text{offered}} \leq B_{\text{offered}}$
(3) The percentage of admitted students who are awarded scholarships ($N_{\%\text{awarded}}$) should be approximately equal to a pre-determined percentage ($N_{\%\text{target}}$): $N_{\%\text{awarded}} \approx N_{\%\text{target}}$
(4) The award amounts must be divisible by \$300 to allow for round hundred-dollar splits across three academic terms.
(5) There is a minimum value for a single scholarship award (S_{min}) but no pre-determined maximum value.

3.6 Optimizing Scholarships

After developing a classifier to predict enrollment, we used the prediction outputs of the classifier as an objective function in optimization. The aim of the optimization was to develop a strategy that disbursed the DNR scholarship budget in a manner that maximized student enrollment. In this work, we used a genetic algorithm (GA) for optimization as GAs are known to work well when we have a well-defined measure to optimize (i.e. student enrollment) but not a well-defined, continuous, and/or differentiable objective function. GAs are also known to find near-optimal solutions quickly, which was essential when we wanted to rapidly outline different budgeting and allocation scenarios early in our modeling.

GAs are a class of evolutionary algorithms and are inspired by biological evolution. GAs generally involve iteratively starting with some population of chromosomes, undergoing selection across this population according to a measure of fitness, using genetic crossover and mutation to produce offspring from the most fit individuals, and then using this offspring as the population for the next iteration [18]. The overall population fitness improves with each iteration and the GA eventually converges towards an optimal solution. As this description of GAs relates to this work, we start with a population of award disbursement strategies whose "genetic material" (chromosomes) are a set of scholarship award values; the measure of fitness to assess these individuals is based on predicted enrollment after accounting for constraints, as detailed below; and the crossover and mutation functions used to create offspring are based on changes of scholarship award values, as described below.

We used the data for the previous year's (2017) admitted class in the optimization of scholarship funds. In all, this was 9,479 students (N_{total}). In this sense, we used data from the year prior to optimize the disbursement for the most recent application year. We pared the data down to a single year's application cohort to avoid having to consider if any of the optimization constraints in Section 3.5 were being violated for each of the application years simultaneously.

We generated a set of possible scholarship awards that spanned S_{min} to a chosen maximum (S_{max}) in \$300 increments and included \$0. Award values had a non-zero minimum value of S_{min} based on historical awards, though stewards of the scholarship fund were amenable to lowering it. Ultimately, as discussed in Section 4.2, this floor was lowered in favor of a scholarship award with a value of $\frac{S_{\text{min}}}{2}$. We did not determine S_{max} beforehand but instead set it such that the optimization procedure did not generate an output that included a S_{max} scholarship award. S_{min} was evenly divisible by \$300 and we generated possible scholarship awards in \$300 increments to satisfy constraint (4) from Section 3.5.

Part of the difficulty of this particular optimization task lies in the fact that awards were to be given in a merit-based manner. As such, the scholarship award for any student is dependent on the awards of students with similar academic profiles. For example, if one was to rank all admitted students in the application pool based on a measure of merit, the minimum possible award given to a particular student would be determined by the award given to the student with the merit that is immediately lower. Similarly, the maximum award a student would be eligible for would be equal to the award given to the student with the merit that is immediately higher. As such, if optimizing on a per-student basis, altering the award for any given student to influence their enrollment decision could result in a cascade during the optimization that subsequently effects every other student's award amount. This results in a very complex fitness landscape when optimizing scholarship awards for thousands of students individually.

As a solution to this issue of an optimization cascade, we first ranked and binned students based on academic merit such that all

students in the same bin received the same scholarship award. To perform this binning, we first sequentially ranked students based on 3 variables: their application academic score, their high school GPA, and their scores on college entrance exams, in that order. This ranking was students' "academic profile." Each student's application academic score was based on an institutional scoring system of their academics and was the primary variable for determining their academic profile. We were provided this metric by the University admissions office - it was not calculated by us. Ties between students having the same application academic score were broken by looking at their high school GPA; any remaining ties thereafter were broken using students' entrance exam scores. Once students were ranked, they were divided into 20 ventiles based on their academic profiles (i.e. students were grouped across every 5th percentile) with each ventile receiving the same scholarship award amount. Using ventiles allowed for us to have sufficient flexibility when exploring the fitness landscape during optimization while also not being so granular as to continually be caught in local extrema. Additionally, ventiles helped mitigate the effect of optimization cascades by giving identical awards to students with similar academic profiles. We refer to each of these ventiles as a "bin" and each bin served as the chromosomal building block for the GA. A single scholarship allocation strategy consisted of the scholarship awards across all 20 scholarship bins and is referred to as an "individual" henceforth when used in the context of the GA. As such, each individual's genetic material can be thought of as being in the form of chromosomes which were composed of scholarship award bins.

We then created a fitness function to evaluate the effect of altering the `reduced_tuition` variable on student enrollment. Specifically, this function took the genetic material of a scholarship individual (i.e. a set of scholarship awards for each bin) and then re-evaluated the `reduced_tuition` variable for each student based on their updated DNR scholarship award. As noted above, we created the `reduced_tuition` variable by taking the tuition due for a student and subtracting their total grants and scholarships; it was the only financial aid and tuition-related variable used in the predictive model. The function re-calculated each student's likelihood for enrollment based on the updated values for `reduced_tuition` using the predictive enrollment model. The final output for the fitness function was a calculation of the number of students predicted to enroll for a given scholarship individual, which we used as the fitness criterion for evaluating individuals.

The organization of the population, individuals, and bins for the GA optimization is shown in Figure 1. We generated an initial population of p individuals by randomly selecting K scholarship awards (one for each bin) and sorting for each individual. For this work, $p = 1000$ and $K = 20$. Each bin effectively contained the same number of students (N_{bin}), which was approximately equal to $\frac{N_{\text{total}}}{K}$. Awards were not unique for each bin and could be duplicated. N_{bin} multiplied by the scholarship award for each bin equalled the funds awarded for that respective bin; the sum of these across all K scholarship bins for a given individual was F_{offered} for that individual. The predicted number of enrollees for each scholarship bin multiplied by the scholarship award for that respective bin equalled the funds spent for that bin; the sum of these across all K scholarship bins for a given individual was F_{spent} for that individual. The number of bins with non-zero award values divided by K was equal to $N_{\%\text{awarded}}$ for an individual.

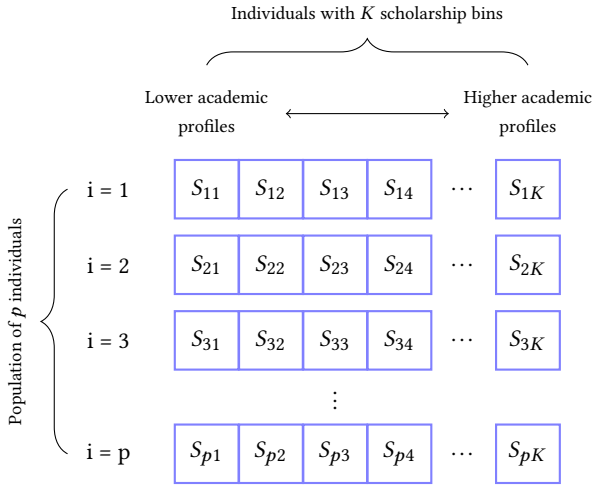

Figure 1: Genetic algorithm setup. Individuals (i) are scholarship allocation strategies of K scholarship bins (j). The population consists of p individuals. Each S_{ij} is a scholarship award value for the i^{th} individual and the j^{th} scholarship bin. The bins are sorted based on academic profile such that $S_{i1} \leq S_{i2} \leq S_{i3} \ldots \leq S_{iK}$ for any given i (but not necessarily across individuals). For this work, $K = 20$ and $p = 1000$.

We penalized each individual's fitness if the optimization constraints in Section 3.5 were violated. We initialized a single penalty coefficient (σ) to 1.0 and then successively enforced each of the following squared penalties for a given scholarship individual:

- if too much was spent on scholarship awards:
  $F_{\text{spent}} > B_{\text{spent}} \rightarrow \sigma = \sigma * (\frac{B_{\text{spent}}}{F_{\text{spent}}})^2$
- if too much was offered in scholarship awards:
  $F_{\text{offered}} > B_{\text{offered}} \rightarrow \sigma = \sigma * (\frac{B_{\text{offered}}}{F_{\text{offered}}})^2$
- if too many students were awarded a scholarship:
  $N_{\%\text{awarded}} > N_{\%\text{target}} \rightarrow \sigma = \sigma * (\frac{N_{\%\text{target}}}{N_{\%\text{awarded}}})^2$
- if too few students were awarded a scholarship:
  $N_{\%\text{awarded}} < N_{\%\text{target}} \rightarrow \sigma = \sigma * (\frac{N_{\%\text{awarded}}}{N_{\%\text{target}}})^2$

Ultimately, we multiplied the output of the fitness function by the penalty coefficient to penalize constraint-violating individuals. If there were no constraints violated, the penalty coefficient remained at 1.0 and the fitness evaluation of the individual remained unchanged. Note that all constraints were given equal weight.

The general process for the GA was as follows. We randomly generated the initial population of individuals as described above. We then calculated the fitness of each individual using the fitness function and took a subset of the most fit individuals from the population (10%) as the basis for the next generation of the population. We employed genetic crossover to this subset of the population to generate offspring. We used two-point genetic crossover, wherein two points were randomly selected along chromosomes and the genetic material from one individual was swapped with that from

another between the two points, much like a two-point crossover mutation in nature. In other words, for a pair of randomly selected individuals, we randomly selected two scholarship bins from ventiles 1 through 20 and all scholarship award values between the two bins from one individual were swapped with those from the other individual and vice versa.

After enough offspring were generated by crossover to refill the population, the offspring underwent mutation. We used three types of mutations: an increase mutation, a decrease mutation, and a swap mutation. For a mutation, we randomly selected an individual and then randomly selected a bin from this individual. The corresponding award for this bin was either increased to another possible award amount (increase mutation), decreased to another possible award amount (decrease mutation), or swapped for another randomly selected award amount (swap mutation). The probability of performing either an increase, decrease, or swap mutation were equal unless the scholarship award value equaled $S_{\min}$ or $S_{\max}$, in which case we eliminated the possibility of a decrease mutation or an increase mutation, respectively. After mutations, we re-sorted the scholarship bins across each individual to ensure students with higher academic profiles received larger awards. We kept the initial subset of the most fit individuals unchanged during crossover and mutation; instead, we altered replicas of these individuals so we could compare the most fit individuals from one generation to those from the next generation. The new generation of individuals then served as the population for the next algorithmic iteration. We repeated the above process for 20 generations of the population and used the most fit individual thereafter as the scholarship allocation strategy. The process for the GA is shown in Process 1.

Process 1: Genetic algorithm process for scholarship allocation (parameters for project are in parentheses)

1: Generate initial population (p = 1000 with K = 20 bins each)
2: Evaluate fitness of each individual (where fitness is enrollment count predicted by XGB classifier)
3: For each of G generations: (G = 20)
4: Keep subset of population with highest fitness (10% kept)
5: Use two-point crossover across individuals to fill population
6: Mutate random bins of random individuals
7: Evaluate fitness of each individual
8: Use individual with highest fitness after G generations

4 RESULTS AND DISCUSSION

Using the methods described in Section 3, we developed a predictive classifier of student enrollment and used it in conjunction with a genetic algorithm that optimizes the allocation of a scholarship fund. Ultimately, the university saw a 23.8% increase in enrollment yield after using our approach. This resulted in millions of dollars of additional annual tuition revenue. The following section presents these results in greater detail in the same order as the methods.

4.1 Predicting Enrollment

Previous studies have shown the effectiveness of ML in predicting enrollment. We examined seven different predictive classifiers for this task. We show the performance of these classifiers in terms of prediction accuracy, AUROC, and F1-score in Table 1. We used the same observations as a test set to compare performance across classifiers; for the test set, the majority class represented 92.8% of observations (i.e. 7.2% of students in the test set eventually enrolled at the University). All classifiers performed similarly in terms of both accuracy and F1-score. Because of the large class imbalance, there were only modest gains in terms of accuracy over the majority class representation. Ensemble classifiers (RF, XGB, and BC) had the highest accuracies while KNN performed on par with the majority class representation (note: it was checked that the KNN model did not predict that all observations were of the majority class). The highest F1-score, meanwhile, was given by the XGB classifier, though it was not substantially higher than other classifiers.

Table 1: Classifier performance sorted by rank across all metrics. Names of classifiers are provided in Section 3.4.

	Model	Accuracy	AUC	F1-score
1.	XGB	93.10%	0.846	0.905
2.	RF	93.06%	0.848	0.901
3.	MLP	93.01%	0.845	0.902
4.	BC	93.05%	0.833	0.901
5.	LR	92.96%	0.805	0.900
6.	SVM	93.00%	0.780	0.900
7.	KNN	92.80%	0.793	0.893

We show ROC curves for the classifiers in Figure 2. The general shape of the ROC curves was similar across the classifiers but with meaningful variation in AUROC. Specifically, RF, XGB, and MLP tended to perform similarly in terms of AUROC and had the highest AUROC values. This is in line with previous work where neural networks tended to perform well when predicting enrollment, even without more complex architectures in this case. That said, the ensemble classifiers performed similarly well for the task at hand.

Demographic data was not used in the models. We expect that including demographic variables in the prediction models would improve predictive performance to some degree, albeit at the expense of potential explicit discrimination with respect to recipient characteristics. As such, we decided to exclude demographic variables when building the classifiers. While doing so limits the degree

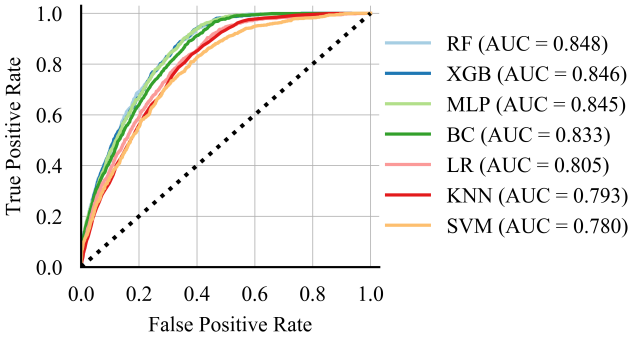

Figure 2: ROC curves for enrollment prediction

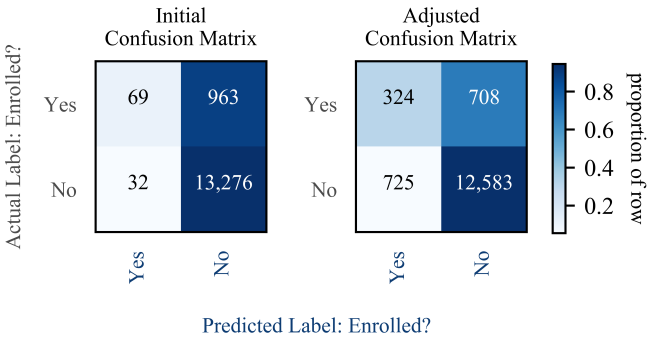

Figure 3: Confusion matrices for predicting enrollment using XGB and a classification threshold of 0.5 (left) and an adjusted classification threshold of 0.22 (right)

of explicit discrimination, the possibility of implicit discrimination remains - particularly with respect to associations between demographics, income, geographic location, and academic performance [21]. Checking and controlling potential demographic imbalances is beyond the scope of this particular work but was handled by stewards of the DNR scholarship fund after optimization.

We examined classifier performance across all metrics and decided to use XGB to optimize scholarship allocation. Prior to optimization, we adjusted the classification threshold for the prediction probability to the nearest one-hundredth such that the number of students predicted to enroll by the model was nearest to the actual enrollment count. By adjusting the threshold in this manner, we used a lower probability decision threshold (0.22) than the value of 0.5 that is typically used in binary classification. We understood that doing so came at the expense of an increased rate of false positives (Type I error) but it also allowed for the prediction counts to be closer to actual counts, which was necessary when discussing predictions with administrative stakeholders. We show the effects of this adjustment in Figure 3, where the confusion matrix using the typical threshold of 0.5 is shown along with the confusion matrix using the adjusted threshold of 0.22.

Of note from the confusion matrices is the degree to which students who were not going to enroll at the University could be predicted while it was much more challenging to identify those who would enroll. This speaks to the selectivity of the University in that many of the candidates who would not enroll were simply those who were not accepted to the University. Concurrently, the difficulty with identifying students who will enroll aligns with the fact that these DNR students are applying to a university that is away from their respective homes and social bases. Also, those that are accepted to the University tend to be of higher academic standing, giving them more potential college choices. Thus, the general likelihood of a DNR student enrolling is difficult to determine when considering potential social factors and college options.

Lowering the classification threshold resulted in predicted enrollment counts in line with what was seen in the data, as shown in Table 2. Lowering the classification threshold also allowed for a greater number of true positives while also balancing the number of false positives and false negatives. We also examined the effect

Table 2: Predicted enrollments after adjusting the classification threshold for test data and all data (training + test data).

	Test Data	All Data
Actual	1,032	5,081
Predicted	1,049	5,166

of similarly adjusting the classification thresholds when using the other ML classifiers and determined that using XGB would still be the most viable for scholarship optimization.

4.2 Optimizing Scholarships

After we developed a model for predicting student enrollment, we used a GA to design a scholarship disbursement strategy. We used the GA in a setup with students grouped in ventiles, with each ventile receiving the same award amount. The genetic material (awards for each ventile) for individuals (allocation strategies) was altered for each iteration of the GA and then fitness was determined. Fitness was based on predicted enrollment after accounting for the violation of constraints. Due to the application review timeline at the University, we did not know which students of the most recent entering class (2018) would be admitted and used the prior year's application data (2017) to develop a disbursement strategy. Because the disbursement strategy relied on students being grouped into ventiles, we easily applied it to the most recent entering class after checking that the two classes were similar. Additionally, the binning strategy and the use of ventiles alleviated concerns about the size of the entering class as specific award amounts were disbursed to proportions of the entering class and not to a fixed count thereof.

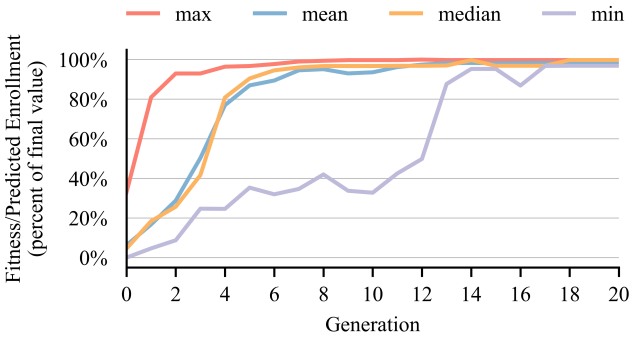

Figure 4: Fitness measures across generations of genetic algorithm. Fitness was equivalent to predicted enrollment.

We show fitness (predicted student enrollment) measures across the population of individuals for each generation of the GA in Figure 4. As expected, the maximum, mean, and median values of fitness increase across generations, though these increases are much smaller for later generations. The minimum fitness values for the population follow a similar trend with some variation. All metrics eventually converge to the predicted enrollment, which is shown as a percentage. Monte Carlo simulations will be used in the future to outline a distribution of likely enrollment counts.

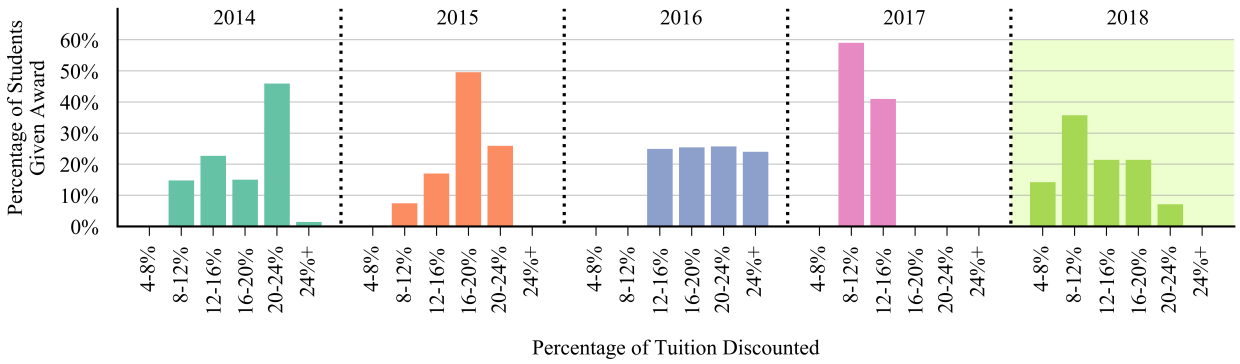

Figure 5: Historical scholarship allocations for the DNR scholarship. The highlighted year (2018) shows the optimized scholarship allocations from this work. Upper bounds for the bins are inclusive. Percentages are of award-receiving students only.

The exact award amounts for the DNR scholarship cannot be disclosed due to University policy. Additionally, the percentage of students receiving scholarship awards was not consistent across previous years. For example, in some years, 30% of accepted DNR students may receive a scholarship while in other years, 70% of accepted DNR students may receive a scholarship. Furthermore, tuition charges change annually at the University. Thus, in an attempt to provide a normalized measure for comparison across entering classes without disclosing exact award amounts, we compare award allocation strategies across time based on the discount on tuition. For example, a student receiving a $5,000 scholarship when tuition is $20,000 receives a 25% discount on tuition. We show previous allocations of the DNR scholarship to scholarship-receiving students as a discount on tuition in Figure 5. This discount on tuition factors in tuition cost for a full-time DNR student but not additional living or educational expenses (i.e. housing, food, books, etc). To further illustrate the use of discount on tuition, when looking at Figure 5, it can be seen that approximately 15% of all scholarship-receiving students received an award that discounted their tuition by 8-12% in 2014 while in 2017, approximately 60% of students received a similar award. It is apparent from examining previous allocations that the manner in which the awards were historically allocated shifted greatly from year to year. As noted previously, these previous allocations were determined by an external consulting service and we could not leverage their underlying approach in this work.

We also show the scholarship allocation strategy for the 2018 entering class (for which the scholarship disbursement was optimized in this project) in Figure 5. This strategy tended to favor smaller scholarships, which aligns with the optimized allocation strategy that Sarafraz et al. reported [22]. In fact, scholarship stewards had initially placed a lower limit on the scholarship awards ($S_{\min}$) during modeling, which was equal to the lowest scholarship amount that had historically been awarded to students. This lower limit was between a 8-12% discount on tuition. After we discussed preliminary results of the optimization and the effectiveness of smaller awards with the scholarship stewards, it was determined that the lower limit on the awards would be changed to $\frac{S_{\min}}{2}$. Thus, the 2018 entering class had some scholarship awards that were lower than those received by previous entering classes. These lower awards discounted tuition by 4-8%. It is also noteworthy that the optimized disbursement strategy gave a distribution of awards that was right-skewed, in contrast to previous allocation strategies, which were predominantly left-skewed or near uniform. This speaks to the idea that smaller scholarships awarded to students of lower merit may be more effective than larger scholarships are for those of higher merit (keeping in mind that students who received smaller awards were also of lower merit for this merit-based scholarship). This aligns with intuition that those with higher academic profiles likely have more college options and require additional recruitment, be it additional financial aid or in some other form.

Table 3: Historical, predicted, and actual yields after scholarship disbursement.

	Timeframe	Yield	% Increase
Historical	2014-2017	10-12%	N/A
Predicted	2018	13.9%	15.8%
Actual	2018	14.8%	23.3%

After we developed the scholarship distribution strategy for the 2018 entering class, the University distributed scholarship awards to admitted DNR freshmen. We then waited as these students indicated their enrollment decisions a few months later. In recent years, the yield for DNR students at the University was about 10-12% with little/no increase, as verified by scholarship stewards, where "yield" refers to the percentage of admitted students who enrolled at the University. Historical yields were not based on an unoptimized or randomized scholarship allocation strategy but were the product of the scholarship allocations derived by an external consulting service. Thus, because we were comparing the results from our approach to those from a previously optimized strategy (and not an unoptimized or random allocation strategy), we expected to see a modest improvement. Instead, we saw an increase in yield that was much higher than our modeling suggested. Table 3 shows the historical yields, the predicted yield based on our optimized approach, and the actual yield based on student enrollment for the 2018 entering class. When comparing to the upper bound on historical yield (12%), we anticipated that the scholarship optimizations

would increase student yield by 15.8% (12% to 13.9%) based on the enrollment figures we had seen during the optimizations. In reality, yield increased by 23.3%. This amounted to hundreds of additional students enrolling with each paying tens of thousands of dollars annually in tuition. Overall, the net effect was an increase in millions of dollars in annual tuition revenue for the University. The University has since incorporated our approach into their enrollment modeling process and will be using it for future disbursements of this scholarship fund. Of note is that the above yields are based on proportions of students that enrolled and the size of the entering class makes little difference when comparing yields. The University also admitted roughly the same percentage of DNR students as years past and nearly all conditions during the application process were identical to previous entering classes. That said, the exact degree to which this increased yield can be causally attributed to the scholarship optimizations warrants further investigation.

5 CONCLUSIONS

In this work, we show how existing data at a university can be used to improve enrollment management. We combine machine learning with numerical optimization and use student application data at a public university to optimize a scholarship fund. We find that the optimized approach increased student enrollment and generated millions in additional tuition revenue. It has since been incorporated into the university's enrollment forecasting.

We show that ensemble classifiers can give strong performance when predicting enrollment and we use a binning strategy based on student merit to make the optimization task more tractable. This strategy eliminated the need for per-student optimizations, thereby limiting the complexity of the fitness landscape during optimization. After optimization, we see that smaller scholarship awards work better for maximizing enrollment. In all, the university had historically seen little/no increase in enrollment yield and we projected that the optimized scholarship disbursement would increase yield by 15.8%. In reality, enrollment yield increased by 23.3%.

Universities are at the forefront of training the next generation of data scientists and developing data-centric tools/techniques. However, they are far behind in applying data science to their own administrative data and processes. This project attempted to move them in this direction. Using a suite of machine learning tools, we were able to increase a university's revenue from a scholarship fund by millions of dollars. We think there are many similar opportunities to harness the power of data science in the realm of education administration, especially in resource allocation.

ACKNOWLEDGMENTS

The authors would like to thank the University data, enrollment, and financial aid stewards for their assistance on this project.

REFERENCES

[1] JK Alhassan and SA Lawal. 2015. Using Data Mining Technique for Scholarship Disbursement. *World Academy of Science, Engineering and Technology, International Journal of Computer and Information Engineering* 2, 7 (2015).
[2] Christopher M Antons and Elliot N Maltz. 2006. Expanding the role of institutional research at small private universities: A case study in enrollment management using data mining. *New directions for institutional research* 2006, 131 (2006), 69–81.
[3] Lin Chang. 2006. Applying data mining to predict college admissions yield: A case study. *New Directions for Institutional Research* 2006, 131 (2006), 53–68.
[4] Michael D Coomes. 2000. The historical roots of enrollment management. *New directions for student services* 2000, 89 (2000), 5–18.
[5] Stephen L DesJardins. 2002. An analytic strategy to assist institutional recruitment and marketing efforts. *Research in Higher education* 43, 5 (2002), 531–553.
[6] Stephen L DesJardins, Dennis A Ahlburg, and Brian P McCall. 2006. An integrated model of application, admission, enrollment, and financial aid. *The Journal of Higher Education* 77, 3 (2006), 381–429.
[7] John Aubrey Douglass. 2010. Higher Education Budgets and the Global Recession: Tracking Varied National Responses and Their Consequences. Research & Occasional Paper Series: CSHE. 4.10. *Center for Studies in Higher Education* (2010).
[8] Jerome H Friedman. 2001. Greedy function approximation: a gradient boosting machine. *Annals of statistics* (2001), 1189–1232.
[9] Cullen F Goenner and Kenton Pauls. 2006. A predictive model of inquiry to enrollment. *Research in Higher education* 47, 8 (2006), 935–956.
[10] Donald E Heller. 1997. Student price response in higher education: An update to Leslie and Brinkman. *The Journal of Higher Education* 68, 6 (1997), 624–659.
[11] John Hood. 1996. The new austerity: University budgets in the 1990s. *Academic Questions* 9, 2 (1996), 82–88.
[12] David S Hopkins. 1981. *Planning models for colleges and universities*. Stanford University Press.
[13] Don Hossler. 2000. The role of financial aid in enrollment management. *New directions for student services* 2000, 89 (2000), 77–90.
[14] Don Hossler. 2009. Enrollment management & the enrollment industry. *College and University* 85, 2 (2009), 2.
[15] Harold A Hovey. 1999. State spending for higher education in the next decade: The battle to sustain current support. (1999).
[16] Larry L Leslie and Paul T Brinkman. 1987. Student price response in higher education: The student demand studies. *The Journal of Higher Education* 58, 2 (1987), 181–204.
[17] Through Educational Data Mining. 2012. Enhancing teaching and learning through educational data mining and learning analytics: An issue brief. In *Proceedings of conference on advanced technology for education*.
[18] Melanie Mitchell. 1998. *An introduction to genetic algorithms*. MIT press.
[19] Ashutosh Nandeshwar and Subodh Chaudhari. 2009. Enrollment prediction models using data mining. *Retrieved January* 10 (2009), 2010.
[20] David Niemi and Elena Gitin. 2012. Using Big Data to Predict Student Dropouts: Technology Affordances for Research.. In *Proceedings of the International Association for Development of the Information Society (IADIS) International Conference on Cognition and Exploratory Learning in Digital Age*.
[21] Dino Pedreshi, Salvatore Ruggieri, and Franco Turini. 2008. Discrimination-aware data mining. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 560–568.
[22] Z Sarafraz, H Sarafraz, M Sayeh, and J Nicklow. 2015. Student Yield Maximization Using Genetic Algorithm on a Predictive Enrollment Neural Network Model. *Procedia Computer Science* 61 (2015), 341–348.
[23] Xanthe Shacklock. 2016. *From bricks to clicks: the potential of data and analytics in higher education*. Higher Education Commission London.
[24] Raj Man Shrestha, Mehmet A Orgun, and Peter Busch. 2016. Offer acceptance prediction of academic placement. *Neural Computing and Applications* 27, 8 (2016), 2351–2368.
[25] Fadzilah Siraj and Mansour Ali Abdoulha. 2009. Uncovering hidden information within university's student enrollment data using data mining. In *Modelling & Simulation, 2009. AMS'09. Third Asia International Conference on*. IEEE, 413–418.
[26] Randy Spaulding and Steven Olswang. 2005. Maximizing enrollment yield through financial aid packaging policies. *Journal of Student Financial Aid* 35, 1 (2005), 3.
[27] Dale Trusheim and Carol Rylee. 2011. Predictive modeling: linking enrollment and budgeting. *Planning for Higher Education* 40, 1 (2011), 12.
[28] Steven Walczak. 1998. Neural network models for a resource allocation problem. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)* 28, 2 (1998), 276–284.
[29] Steven Walczak and Terry Sincich. 1999. A comparative analysis of regression and neural networks for university admissions. *Information Sciences* 119, 1-2 (1999), 1–20.
[30] Darrell M West. 2012. Big data for education: Data mining, data analytics, and web dashboards. (2012).