PLOS ONE

RESEARCH ARTICLE

Predicting time to graduation at a large enrollment American university

John M. Aiken[1,3]*, **Riccardo De Bin**[2], **Morten Hjorth-Jensen**[1,3,4], **Marcos D. Caballero**[1,3,5]

1 Department of Physics, Centre for Computing in Science Education, University of Oslo, Blindern, Oslo, Norway, **2** Department of Mathematics and Statistics, University of Oslo, Blindern, Oslo, Norway, **3** Department of Physics and Astronomy, Michigan State University, East Lansing, Michigan, United States of America, **4** National Superconducting Cyclotron Laboratory and Facility for Rare Ion Beams, Michigan State University, East Lansing, Michigan, United States of America, **5** CREATE for STEM Institute, Michigan State University, East Lansing, Michigan, United States of America

* j.m.aiken@fys.uio.no

Citation: Aiken JM, De Bin R, Hjorth-Jensen M, Caballero MD (2020) Predicting time to graduation at a large enrollment American university. PLoS ONE 15(11): e0242334. https://doi.org/10.1371/journal.pone.0242334

Editor: Gábor Vattay, Eötvös Loránd University, HUNGARY

Received: May 13, 2020

Accepted: October 30, 2020

Published: November 13, 2020

Data Availability Statement: All relevant data are available from Zenodo (DOI: 10.5281/zenodo.4114005).

Funding: MDC - The Michigan State University College of Natural Sciences including the STEM Gateway Fellowship and the Lappan-Phillips Foundation https://msu.edu. MDC - the Association of American Universities https://www.aau.edu/. MDC, MJH - the Norwegian Agency for Quality Assurance in Education (NOKUT), which supports the Center for Computing in Science Education. https://www.nokut.no/en/. MJH - INTPART project

Abstract

The time it takes a student to graduate with a university degree is mitigated by a variety of factors such as their background, the academic performance at university, and their integration into the social communities of the university they attend. Different universities have different populations, student services, instruction styles, and degree programs, however, they all collect institutional data. This study presents data for 160,933 students attending a large American research university. The data includes performance, enrollment, demographics, and preparation features. Discrete time hazard models for the time-to-graduation are presented in the context of Tinto's Theory of Drop Out. Additionally, a novel machine learning method: gradient boosted trees, is applied and compared to the typical maximum likelihood method. We demonstrate that enrollment factors (such as changing a major) lead to greater increases in model predictive performance of when a student graduates than performance factors (such as grades) or preparation (such as high school GPA).

1 Introduction

University students must meet a number of objectives to obtain degrees and in many cases this can prolong their time at the university [1] or they drop out altogether [2]. During their studies, American students may take on substantial financial obligations when choosing to pursue degrees and extending beyond the "four-year degree" can greatly increase the cost of obtaining that degree [3]. Thus understanding the paths that students take towards degree completion can help faculty and administrators better serve student populations to meet their educational goals. Tinto's Theory of Drop Out [4] has seen large acceptance in its ability to describe the factors that influence a student's path towards degree completion [5–9]. Tinto theorized that a student's college drop out decision is mediated by two conglomerate features: 1) educational goal commitment, and 2) institutional commitment. He further theorized that these commitments are dynamic. Students begin their university studies with initial commitments that are

of the Research Council of Norway (Grant No. 288125). https://www.forskningsradet.no/en/call-for-proposals/2019/funding-for-international-partnerships/. MJH - U.S. National Science Foundation (Grant No. PHY-1404159). https://www.nsf.gov/index.jsp. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

then mediated by how students participate in the academic and social systems of the university. Tinto suggested that using institutional data such as that collected by university registrars, could quantify these relationships and provide predictive models for determining student's at-risk to drop out or alternatively, graduate.

This paper examines student ultimate success at university. We define this success as obtaining a bachelor's degree. It uses 20 years of institutional data for 160,933students attending a large enrollment American research university. This paper examines two specific research questions:

1. In the context of Tinto's Theory of Drop Out [4], what factors (such as grades, participation in a major, student background, etc.) contribute to the time it takes to obtain a degree at a large enrollment research university in the United States? How does the contribution of these factors change for longer durations until graduation?
2. How can recent innovations in statistics and machine learning, such as gradient boosting and xgboost, improve educational model performance?

We focus on a comparison of student participation in an academic system, their involvement in the social system, and their initial conditions due to high school training. We demonstrate that enrollment factors (e.g., changing a major) and academic performance are more important to predicting when a student graduates than pre-college experiences (e.g., high school GPA). This study focuses on student educational commitment using registrar data. Additionally, we compare traditional statistical modeling (maximum likelihood estimation) to new techniques from machine learning (gradient boosting [10]) and demonstrate that the machine learning methods are more effective at estimating the function predicting time-to-graduation.

It is important to use an analysis technique that respects the dynamics assumptions of Tinto's Theory of Drop Out. In this paper we use a Discrete Time Hazard Model framework [11]. Discrete time hazard model is useful when the classic Cox regression model assumption that events happen on a continuous interval no longer hold valid [11]. This is true when data has highly discretized time intervals such as semesters enrolled. Traditionally this modeling is done with logistic regression via maximum likelihood estimation [11]. In this paper we compare two models that fit the logistic model including logistic regression (maximum likelihood estimation) and a gradient boosted tree model (xgboost, [10]).

2 Background

In this section Tinto's theory [4] and the use of that theory will be described, the use of discrete time hazard model [11] to predict when students graduate and/or drop out will be described, and the value of a new machine learning method known as gradient boosting will be described [10]. The functional descriptions of the discrete time hazard model and gradient boosting are described in Section 4.

2.1 Tinto's theory of drop out

Tinto's Theory of Drop Out describes a students intent to drop out based on the interplay of two quantities: 1) a student's commitment to education, and 2) a student's commitment to a specific institution. A student's commitment to education is mediated by the initial state of a student entering the university and the dynamics that occur while the student attends university. These dynamics are dictated by a student's participation and acceptance into the social and academic communities that are at a university. A student's commitment to an institution is tempered by many factors such as the educational goals available at an institution (e.g., a

technical university degree offerings versus a liberal arts university), family commitment to a university, and social acceptance at the university. While Tinto's theory explicitly attempts to describe why students drop out from college, it is not uncommon that this theory is used to study student graduation from college given that graduation is an alternative outcome to dropping out [1, 9, 12]. In this paper we follow this trend to extend beyond the original goal of Tinto's drop out model by using it as a framework to predict the time it takes a student to graduate. In this paper we focus on a student's commitment to education as the framing for features that predict when a student will graduate if they do so.

It is common to characterize Tinto's theory of dropout as a discrete time hazard problem [1, 8, 9, 12, 13]. Tinto's theory posits that a student's educational commitment changes over time as they work toward's a degree and that effects, such as high school GPA, that may be strong in the beginning of a degree program are weak toward's the end of a degree program. Discrete time hazard modeling provides a systematic method for examining the various effects on the probability to graduate over time [11]. To that end, discrete time hazard modeling has been used to examine the dynamic impact of various effects that Tinto predicts effect a student's educational commitment. College GPA has been demonstrated to have a profound but diminishing time varying effect on graduation [1, 8, 14]. Financial aid and money spent on student services has also demonstrated a time varying (diminishing) effect on the probability to graduate [1, 14]. First generation student's are considered high risk for drop out but this effect diminishes over time indicating that first generation students need specific resources other students may not [9]. Non-traditional enrollment factors such as delaying enrollment, working while enrolled, and stopping education for some period of time can all have a negative time varying effect on graduation [12]. In each case these studies showed a statistically significant time varying impact that supports the dynamic claims of Tinto's theory.

A student's preparation has long been known to impact a student's ability to graduate. Preparation is assessed in many ways and can be represented as the experiences a student has in school and also their present innate ability to perform a specific function. Math preparation correlates with both performance in university and graduating [15, 16]. The same is true for physics and english preparation [17–19]. High school GPA and SAT scores typically account for some but not all of student success at university (GPA, etc.) [20–22]. Preparation for university often can be experienced differentially as well. Women in STEM (Science, Technology, Engineering, and Mathematics) degree programs report having less access to laboratory experiences in high school, are encouraged towards science by their father's differently, and overall have a different preparation than their male counterparts [19, 23].

A student's demographics can include the student's gender, race, the financial support they can expect from their family, and if a student is the first in their family to attend college. Race has long been shown to be a factor in whether a student graduates or not. Black students are less likely to graduate than fellow white students when adjusted for socio-economic status and academic ability and they are more likely to have unwelcoming experiences in STEM programs while at university [1, 23, 24]. Female students are also likely to have unwelcoming experiences that cause them to switch from STEM programs [23]. However they have been shown to be equally likely to graduate or more likely to graduate in comparison to their male counterparts [23, 24]. The financial support both in terms of loans and scholarships and the socio-economic status of a student's family have long been known to be a factor in university graduation with students who have more financial support typically being more likely to graduate [1, 24–27]. First generation students are less likely to participate in extra-curricular activities at university [28, 29] and are more likely to dropout even when adjusting for race, family income, gender, and preparation [9]. Ultimately, American students from different backgrounds often see different success rates at university

due to the different experience they have at the university due to their race, gender, or socio-economic status.

2.2 Gradient boosting

In this paper we use a novel method known as "gradient boosting". Gradient boosting can solve a large number of statistical modeling problems including logistic problems as found in this paper. Logistic problems are a group of statistical problems that assume that a sigmoidal function made from parameters θ and data $\boldsymbol{x}$ (i.e., $[1 + e^{-\theta^T \boldsymbol{x}}]^{-1}$) approximates the probability $P(Y|X)$. In plain language, the model uses input data X to determine how likely the outcome, Y, is to be true. In the context of this paper that would mean the input data X is used to predict whether a student will graduate in the following semester. In education research, the maximum likelihood method has commonly been used to find the solutions to logistic problems. Maximum likelihood solves the logistic regression problem by picking parameters θ that maximize the log likelihood function [30].

Since there is no closed form solution to this likelihood equation the solution is found iteratively via some optimization algorithm [30]. Maximum likelihood is the typical default solution for logistic problems in most statistical packages such as statsmodels [31] in python and glm in R [32].

Gradient boosting produces the solution to the logistic problem in a different way. The log-likelihood is maximized (or, in machine learning terms, the negative log-likelihood is minimized) in small steps: at each step, called boosting iteration, the gradient of the log-likelihood is computed to identify the best direction for the maximization, and it is then fed to a base learner, through which the model is updated. Different choices of base learners (in this paper we will use trees) provide different solutions, guaranteeing maximum flexibility [33]. An early stop, through a tuning parameter which controls the number of iterations, prevents overfitting and provide a better bias-variance trade-off [34]. If the base learner is weak enough in comparison to the signal-to-noise, it can be shown that, at least for a continuous response, boosted models outperform their unboosted versions in term of MSE [35]. Several modifications of the boosting algorithm have been proposed, including stochastic gradient boosting, that uses column and row subsampling to further avoid overfitting and even better deal with the bias-variance trade-off issue [36].

Gradient boosting has been demonstrated to produce better fit models in comparison to traditional methods across a number of domains both in binary classification and in hazard modeling [37–39]. Boosted logistic models have been demonstrated to be more effective at fitting data than models that use maximum likelihood estimation [40]. The structure of educational data can be highly complex often containing many sub groups that have different effects [41]. It is likely that gradient boosting can provide better parameter estimation for education research questions as well.

In this paper we use the Extreme Gradient Boosting (xgboost) algorithm [10]. Xgboost implements stochastic gradient boosting [42] with column and row-wise subsampling, regularization, decision tree base learners with a custom tree split finding algorithm, and an in-model data imputation system for missing data. The specifics of xgboost are explained in Section 4.

3 Data set

The data in this study comes from registrar information from a large enrollment American research university [21]. This institution is a highly ranked, mid-western, large enrollment university with approximately 50,000 students enrolled annually and a graduation rate of 78%. It

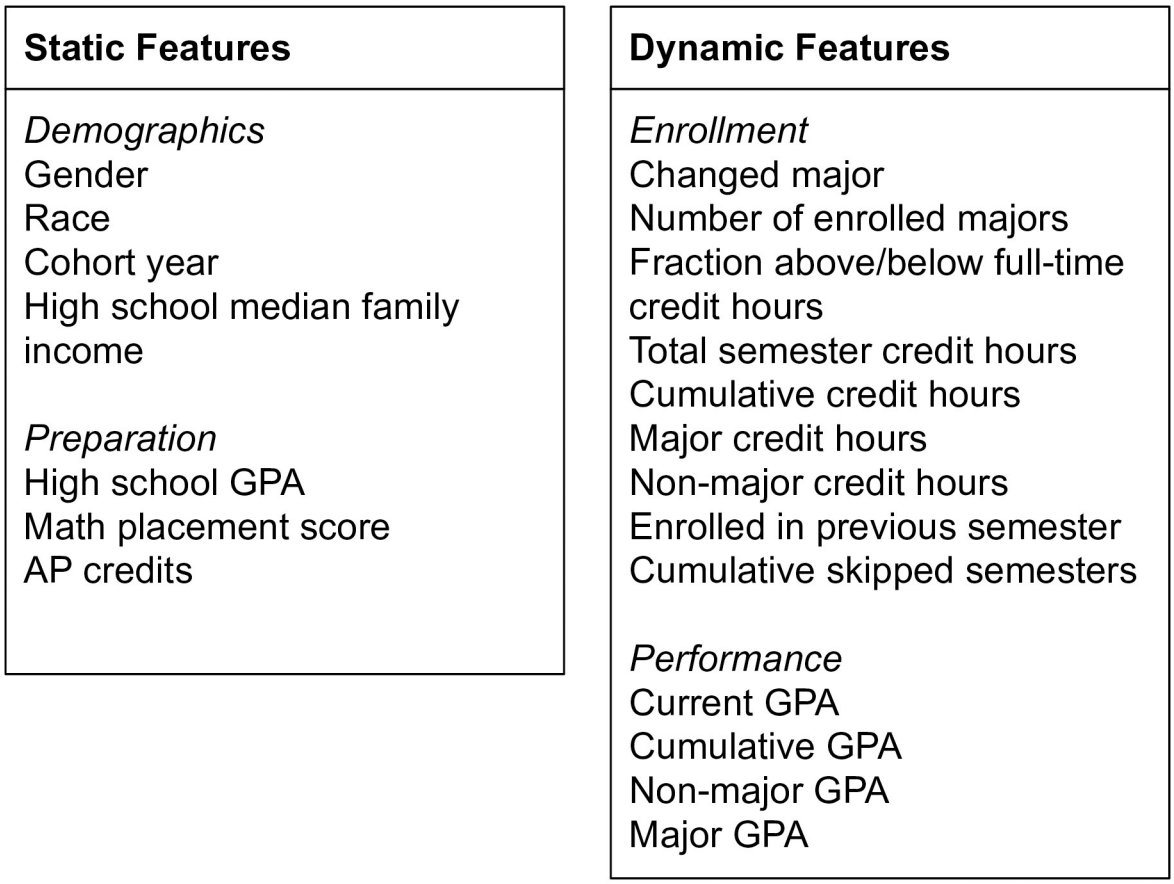

Fig 1. Data model for all models presented in this paper.

https://doi.org/10.1371/journal.pone.0242334.g001

includes timestamped course grades, demographics, majors declared, degrees awarded, and preparation information for 160,933students for the years 1992 to 2012. Student's are included in the study if they remain enrolled for at least 5 semesters. The 5 semester mark is chosen because a large number of students(18.3%) remain enrolled for only four or less semesters taking core courses. These students do not graduate the university and may transfer elsewhere as their institutional commitment may be low [4]. This is aligned with other research which has shown that students are most likely to drop out in the first year [1, 43, 44]. 88.02% students who remain enrolled for at least 5 semesters graduate from this university. The dynamics of the first four semesters is confirmed by applying the models presented in this paper to those semesters. Across the first four semesters neither of the models is able to produce an F1-score above 0.1 thus having a very poor fit and a very low confidence in any interpretation that could be produced during these semesters.

In this study, data is organized by semester into of four categories: demographics, preparation, enrollment, and performance (Fig 1). Demographics and preparation are considered "static" features. That is, they are not changing over the course of the study. The enrollment and performance features are considered "dynamic". That is, they are cumulative and values can change as students progress in their studies.

The demographics data includes gender, race, cohort year, and the median family income for the zipcode the high school the student attended. The gender category is binary: male or female. The race category uses the IPEDs definitions [45] which is then reduced to a binary

category of white/asian or other. This is due to two reasons: 1) the university has a primarily white/asian population, and 2) in this paper we establish model comparison's via out-of-sample prediction accuracy. Thus due to the small minority population, out-of-sample, per semester data may not include all racial groups with sufficient numbers (>100, [46]). The cohort year is the year the student begins their first courses. The median income comes from the 2011 American Community Survey 5-Year Estimate [47]. If the student has a reported high school GPA, then we also know the zip code of the high school the student attended. This zipcode is matched to the 2011 census data zip codes. The census data includes the median incomes for families living in that zipcode. The log of the median income is then recorded per student. If the zip code data is missing the median income is imputed as described in Section 4. This method is not an attempt to provide a course grained estimate of the socio-economic status of the student. Instead, it is a method to obtain a measure of the socio-economic status of the high school the student attended. Student learning can be impacted not only by a student's personal socio-economic status, but also the socio-economic status of the learning community they belong to [48]. We do not have individual level socio-economic data such as parental income. Nor do we have data on financial aid status.

The preparation data includes the per student reported high school GPA, a math placement score, and if amount of Advanced Placement credits if any. The math placement score comes from a 30 item test that the university requires students to take upon admission. This test has been used for the entirety of the study. A higher score will place the student in a higher math course up to calculus 1. High school GPA is also reported for each student. In both the cases of high school GPA and math placement score the data can be missing for several reasons. High school GPA is not required to be reported by applicants to this university, thus high school GPA is not always recorded for each student. If students have transfer courses for higher math courses (e.g., calculus 1) then the student does not need to take a placement test to determine if they can begin in calculus 1. In the case of missing data the math placement score and the high school GPA are imputed as described in Section 4.

The enrollment features are organized per semester. They include whether a student changed their major in that semester, the number of currently enrolled majors, the ratio of credit hours to a full time load (12 credit hours), the total registered semester credit hours, the total cumulative credit hours accrued, the non-major credits the student has registered for in this semester, the major credits hours enrolled per semester, whether the student was enrolled in the previous semester, and the cumulative number of skipped semesters.

The performance features are organized per semester. They include the current semester GPA, the cumulative GPA, the current semester's GPA for courses outside of the student's current major, and the current semester's GPA for courses within the student's current major.

4 Methods

In this paper we have used a discrete time hazard modeling framework for all models that are presented. We begin by comparing the logistic regression model to a gradient boosted (xgboost) [10] model to produce the most predictive model. In this section we will describe the discrete time hazard modeling framework, the logistic regression equation, the xgboost model equation, and the statistics and evaluation methods that we have used to determine the xgboost model is the best predictive model. We will then describe the methods we used to evaluate what features the model uses to produce predictions. The superior xgboost model is then used to describe what factors are most predictive of students graduating in a particular semester.

Michigan state university

MSU Study ID: STUDY00000043

Principal Investigator: Marcos Daniel Caballero
Category: Exempt 1
Exempt Determination Date: 1/8/2018
Title: Studying student pathways in College STEM curriculum and the factors that support or inhibit success This project has been determined to be exempt under 45 CFR 46.101(b) 1. There is no consent necessary because it is considered exempt.

4.1 Discrete time hazard models

Discrete time hazard models are, essentially, logistic regression (or other classification) models calculated per time step for time changing data. They control for time to event predictions when the times to events are simultaneous or explicitly discrete and there is new data being collected over the course of the study period [11]. Traditional hazard analysis such as Cox regression is not capable of doing this type of analysis due to the duration being measured being a common value amongst many students [11]. This is due to the likelihood function of a continuous event duration model assumes independent and unique durations (i.e., the time it takes to graduate). When these durations are not unique, they can lead to overfitting. In this study the time unit used is a semester [1]. Thus the models attempt to predict whether a student will graduate in the immediate following semester. If a student drops out from the university, they are not included in following semester data set.

4.2 Logistic regression

The logistic regression equation used in this paper is as follows:

$$\log\frac{y(t)}{1-y(t)} = \beta_0(t) + \boldsymbol{\beta}_S^T \boldsymbol{X}_S + \boldsymbol{\beta}_D(t)^T \boldsymbol{X}_D(t) \tag{1}$$

Where $y(t)$ is the likelihood to graduate in the following semester and $\boldsymbol{\beta}_S$ and $\boldsymbol{X}_S$ are static features indicated by subtext S (Fig 1). The dynamic terms (indicated by subtext D) are calculated per semester t from semester 5 until semester 16 [1]. ϵ is then the irreducible random error. This is effectively an iteratively calculated multinomial logistic regression model [1]. The model is fit using the maximum likelihood estimation method. No regularization is used.

The logistic regression model is built using the statsmodels library in python [31].

4.3 Gradient boosted trees

Xgboost is an implementation of a stochastic gradient boosting machine [10, 42, 49] that can also be used to attempt to solve the logistic equation. Gradient boosted machines can be seen as models that are iteratively fit on the current residuals starting from the null model. The additive collection of these models produce the output. Xgboost uses decision trees as its base learners.

The gradient boosted model is thus:

$$\log\frac{y(t)}{1-y(t)} = F_0 + \sum_m^M h_m(\boldsymbol{X}(t)) F_m(\boldsymbol{X}(t)) \tag{2}$$

Where m is the iteration index, $h_m(X)$ is the previous iterations residual model and $F_m(X)$ is the current iteration's model fit to previous iteration's residuals. The total number of iterations is set by M.

A gradient boosted machine can use any learner for the iterative procedure. In this paper we use the decision tree learner as implemented in Xgboost [10]. Decision trees are models

that use a tree like structure to fit data and produce regressions and classifications. For each "leaf" node in a tree, a specific decision is made that separates data into two diverging paths. Each node represents a single variable. Categorical variables (e.g., gender) are simply split by the category. Continuous variables (e.g., semester GPA) are split by a decision boundary (e.g., GPA > 3.0). This boundary is typically determined through iteratively fitting the model to find the best boundary.

In Xgboost, each decision tree is trained from a randomly sampled set of rows and columns. Each tree is grown up to a maximum depth using a leaf growing algorithm that estimates whether an additional leaf will produce a better or worse tree [10]. Thus there is no separation between static and dynamic features as in the logistic regression model for each semester. All xgboost models are built using the xgboost library in python [10].

Xgboost has a built in algorithm to deal with missing data [10]. When missing, data is imputed for each tree by selecting the most common decision path for a node. Because a feature may appear in many trees in the model, the decision boundary can be very different per tree in comparison to an mean imputation scheme.

Xgboost models have hyperparameters that define how the model is to be constructed prior to the model being trained. These hyperparameters govern two classes of model design: 1) how the boosting functions, and 2) how the trees are grown and structured. In the case of boosting, the most common hyperparameter used is the number of total boosting learners. For the tree structure this includes the maximum depth a tree can grow to and the fraction of variable and row sub-selection. Typically these hyperparameters are determined by grid search when the total combination of hyperparameters is below 1000 combinations [50, 51].

Because the learners in the xgboost model are decision trees and not linear models, the xgboost model does not report typical coefficient values like in a traditional logistic regression. Instead they report feature "importances" known as "gain". Each time a variable is used in a tree, the tree is built optimally by splitting in the optimum location. The increase in accuracy due to this split is the gain. The feature importance for a specific variable reported then is the average gain across all the instances that the variable is used in in the model. Xgboost uses a custom split finding algorithm that compares the utility of growing a tree to its maximum depth based on increase in accuracy [10].

4.4 Model evaluation

The goal of this study is to produce a predictive model of when student's graduate. We use a number of techniques to verify the veracity of the model and increase its accuracy (Fig 2). These include: splitting the data into a training/testing sets, imputing missing data, and picking custom thresholds for the predicted probability of graduation. We also limit the over-estimation of a variable's impact on when a student graduates by weighting explanation methods with the F_1-score [30]. The following section will explain these techniques in detail.

The data processing and model evaluation follows the procedure shown in the flowchart in Fig 2. Data for the discrete time hazard model models are:

1. queried from the pathways database [21]
2. Continuous values such as high school GPA and semester GPA are rescaled using the z-score. The means and standard deviations are determined by either the starting year cohort (high school GPA) or the timestamp for the enrolled semester (semester GPA). The median income is scaled using the logarithm.

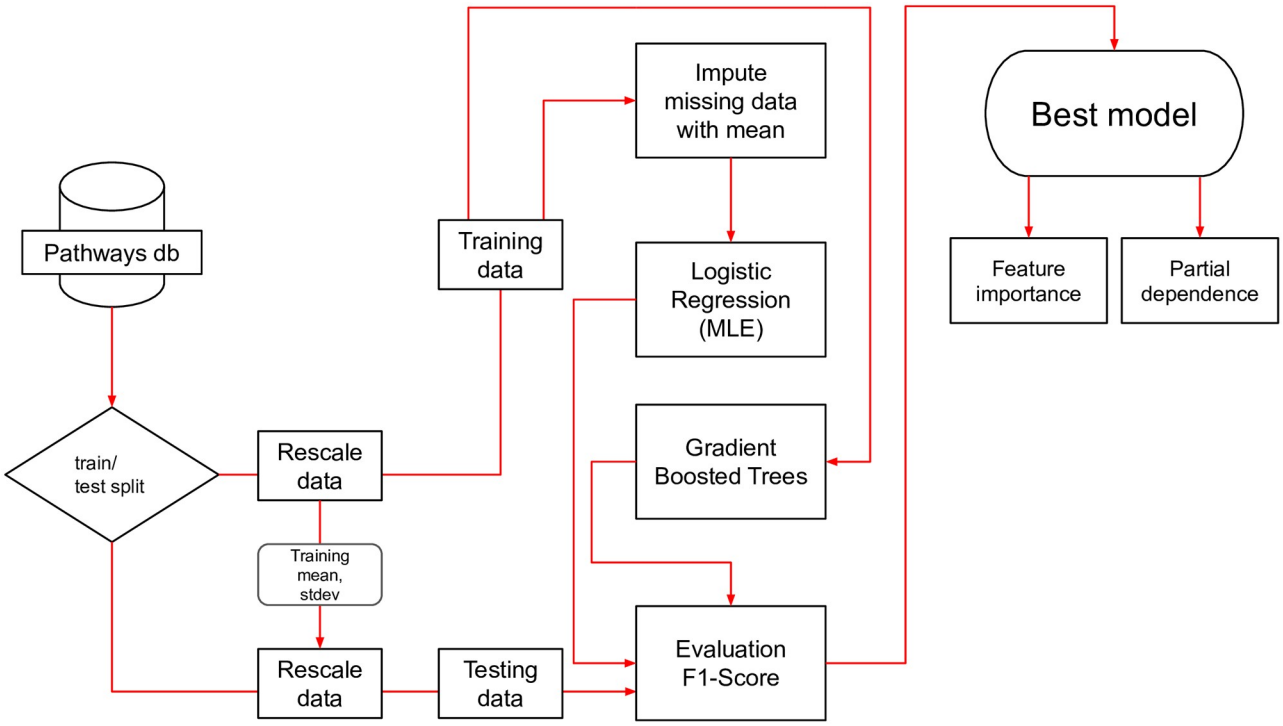

Fig 2. Model evaluation flowchart. Data is split into testing and training sets and is evaluated for two separate models: logistic regression solved by maximum likelihood and gradient boosted trees (xgboost). Missing data for the xgboost model is not imputed from the mean and instead uses the built in imputation engine within xgboost [10].

https://doi.org/10.1371/journal.pone.0242334.g002

3. Data is split into training (50%) and testing (50%) sets. Since the discrete time hazard models are evaluated per semester data is split by student and not by semester. Thus a student in training data in semester 5 is also in training data for subsequent semesters.
4. Missing data that is input into the logistic regression model is imputed from the means calculated during the rescaling step. Missing data for the xgboost model is not imputed prior to fitting the model since the xgboost algorithm has an imputation engine built in.
5. Each model is trained and evaluated using the F_1 score. In-sample F_1 scores are used to determine the best threshold for splitting the predicted probability distributions for classification (i.e., does the student gradute in the next semester).
6. Models using the selected threshold are then evaluated using out-of-sample test data to compute the F_1 score.

Several data handling procedures were used to prevent model overfitting, data leakage, and poor predictive performance due to unbalanced class issues. To evaluate the predictive capability of each model, data in this paper is separated into training and test sets for each semester that the discrete time hazard model model is applied. Separating data into partitioned training and testing sets is important to prevent a too optimistic evaluation of the model error [30]. Without having hold out data to evaluate the model performance, we have no way of knowing how a model will predict the outcome of data it has not seen before [30, 46, 52]. Using hold-out data is atypical in recommendations within education research for assessing predictive ability [53, 54] and is not used typically in modern papers using discrete time hazard analysis (e.g., [1]).

Data leakage could occur between two semesters if, for example, a student in semester 6 is in a training set and the same student is in a test set in semester 7 [46]. This is due to some of the features being cumulative thus they would carry forward information from the previous training period into the next test period. Thus, students are identified prior to model training as belonging to the training or testing data set. Students in the test data set are only given to the model to evaluate model predictive performance.

Students do not always have an entry in the database for their high school GPA or their math placement score. This can be due to a variety of reasons such as the high school GPA was not reported, the reported high school GPA was self reported and potentially untrustworthy, and the math placement score was not given to the student because they transferred math credits from another institution. To prevent errors associated with removing data [55], this data is imputed in one of two ways. For the logistic model data is imputed from the mean for the student's cohort year. For the xgboost model data is imputed within the model. Xgboost will actively impute missing data within each tree learner based on the most likely branch for each decision node (see Section 4 and [10] for more details).

The predictive ability of models in this paper is estimated using the F_1-score and the recall [56]. The F_1-score is the harmonic mean of the precision and recall and is calculated with the following equation:

$$F_1\text{-score} = 2\frac{\mathrm{Pr} \cdot \mathrm{Re}}{\mathrm{Pr} + \mathrm{Re}} \tag{3}$$

$$Pr = \frac{\text{true positive}}{\text{true positive} + \text{false positive}} \tag{4}$$

$$Re = \frac{\text{true positive}}{\text{true positive} + \text{false negative}} \tag{5}$$

The F_1 score is over the range 0 to 1 with 1 being perfect performance and 0 being random performance. Precision is defined as the ratio of true positives to the sum of true positives and false positives. That is, it is the ratio of true predicted graduations per semester to the sum of the true predicted graduations and those predicted to graduate but actually do not. Precision is over the range 0 to 1 with 1 being perfect performance and 0 being random performance. Recall is defined as the ratio of true positives to the sum of true positives and false negatives. That is, it is the ration of the number of students predicted to graduate to the total number of people who actually graduated per semester. Recall is over the range 0 to 1 with 1 being perfect performance and 0 being random performance.

We use the F_1-score over other statistics (such as the area under the reciever operating curve (AUC), [30]) because it balances the trade off between high precision (which includes falsely labeling students as graduated) and high recall (which includes falsely labeling students as not graduating). We use the Recall score in the cases of understanding model performance for sub-groups (such as under-represented minorities) because it gives a more direct interpretation of how accurately the model selects the graduating case.

Students are enrolled for many semesters however they only graduate in a single semester. Thus the majority class in each semester is that a student will *not* graduate in the following semester. This unbalanced class issue can lead to under-fitting of the model specifically such that models simply predict the majority class (in this case not graduating) far too often [57]. We attempted to use an over-sampling technique [58] to address this issue however this did not increase model performance. Instead we choose custom thresholds for the model

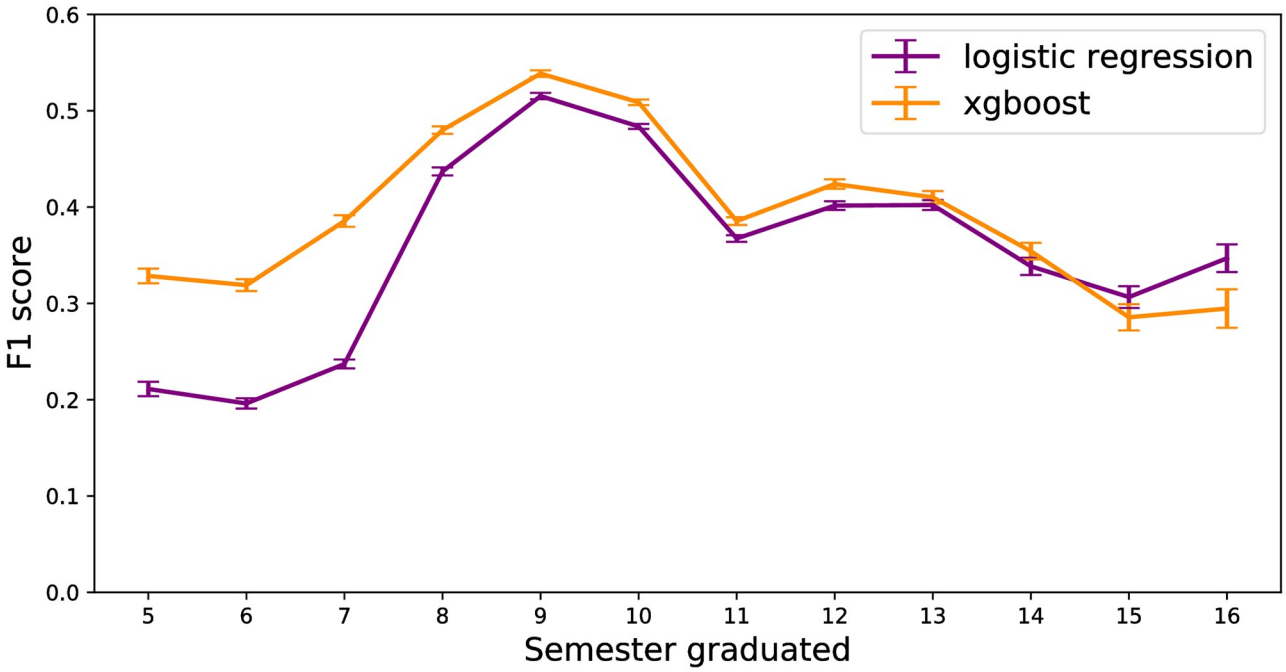

Fig 3. Out-of-sample F_1 scores for all models per semester. Xgboost clearly increases the precision and recall tradeoff for the bulk of semesters. Error bars are the bootstrapped standard deviation of the F_1 score.

https://doi.org/10.1371/journal.pone.0242334.g003

probability distributions for predicting when a student graduated. The custom thresholds are picked using the F_1 scores calculated from the training data [59]. The best threshold is calculated via a grid-search between a threshold of 0 to 1 using increments of 0.01. This still, typically, does not increase F_1 scores.

Beyond predicting when a student will graduate we are interested in the factors that explain why a student was predicted to graduate (or not graduate). A typical logistic regression model using maximum likelihood produces coefficients that represent the magnitude and direction a particular feature given the other variables. For gradient boosted trees these values are calculated differently. Xgboost uses instead the average gain in prediction accuracy across all instances a feature is used in each tree learner. This is commonly called the "feature importance". Additionally we can estimate how much a feature, depending on individual values, contributes to the predicted probability of graduating. In this case we use a method called "partial dependence" that was originally developed to use with gradient boosted models [49]. Both the xgboost model and the logistic regression model have a different level of confidence in the prediction per semester (Fig 3). Due to this variable confidence our confidence in feature importances and partial dependences varies as well. In order to account for this confidence we have weighted the feature importances and partial dependences with the out-of-sample F_1 score. This is explained below.

We estimate the effects of model features on prediction using the using the weighted mean of the feature importances. For all 11 semesters the feature importance is weighted by the overall F_1-score for the semester following this equation:

$$\boldsymbol{\beta}_{weighted} = \sum_{i=0}^{N=11} F_1^T \boldsymbol{\beta}_i \tag{6}$$

This then allows for the cumulative weighted features in Fig 4 and the weighted features in Fig 5.

We also estimate the effects of model features on prediction using the partial dependence [49]. A partial dependence plot represents the average model output for a single variable (S) across the entire feature space (C) in the context of all other features [49]. This means that for a specific feature S, the exact values of S are first, fixed for all rows in the data set for each unique value of S. Then the model is calculated per value. The average contribution of a specific unique value to the overall predicted probability of graduation is then considered the partial dependence. Partial dependence is then assumed to be a continuous function over the entire feature space of S. Partial dependence is estimated using the following equation as:

$$\hat{f}_{x_S}(x_s) = \frac{1}{n}\sum_{i=1}^{n}\hat{f}(x_S, x_C^i) \tag{7}$$

The partial dependence is calculated for each semester and is weighted by the F_1-score for that semester (Figs 6 and 7) in a similar fashion to Eq 6.

Partial dependence allows the researcher to see the direct contribution to the predicted probability of a given variable value. Because variables often have many values, they can give fluctuating contributions to the predicted probability. In linear models, it is assumed that variables produce linear contributions to the partial dependence [30, 49] and thus coefficients of a linear model represent unit increases [30, 60]. For nonlinear models such as gradient boosted trees, this assumption is relaxed and variables can produce nonlinear contributions to the predicted probability.

5 Results

In this study we have two broad research questions: 1) exploring the effectiveness of a gradient boosted logistic model in comparison to traditionally solved logistic models, and 2) quantifying the components of Tinto's Theory of Drop Out in the context of who does or does not graduate at this university during the study period. In this study we have two broad research questions: 1) quantifying the components of Tinto's Theory of Drop Out in the context of who does or does not graduate at this university during the study period, and 2)exploring the effectiveness of a gradient boosted logistic model in comparison to traditionally solved logistic models. In this section we will first describe the effectiveness of the gradient boosted model in comparison to the traditional maximum likelihood model. Then we will describe the results that are drawn from the gradient boosted model in the context of Tinto's theory. In this section we will first describe the results that are drawn from the gradient boosted model in the context of Tinto's theory. Then we will describe the effectiveness of the gradient boosted model in comparison to the traditional maximum likelihood model.

5.1 Gradient boosting

The models in this study attempt to predict whether a student will graduate or not during the observation period of being enrolled for 5-16 semesters. The xgboost model is generally more effective than the logistic regression model except in the last semesters (15-16) studied as assessed by the out-of-sample F_1-score (Fig 3). Xgboost is particularly more effective in highly imbalanced case of students graduating (approximately 5% per semester (Fig 8)) within $\leq$8 semesters (Fig 3). In the more balanced case (semesters 8-14 have approximately 20% of the students eligible to graduate do so (Fig 8)), Xgboost still performs better than the logistic

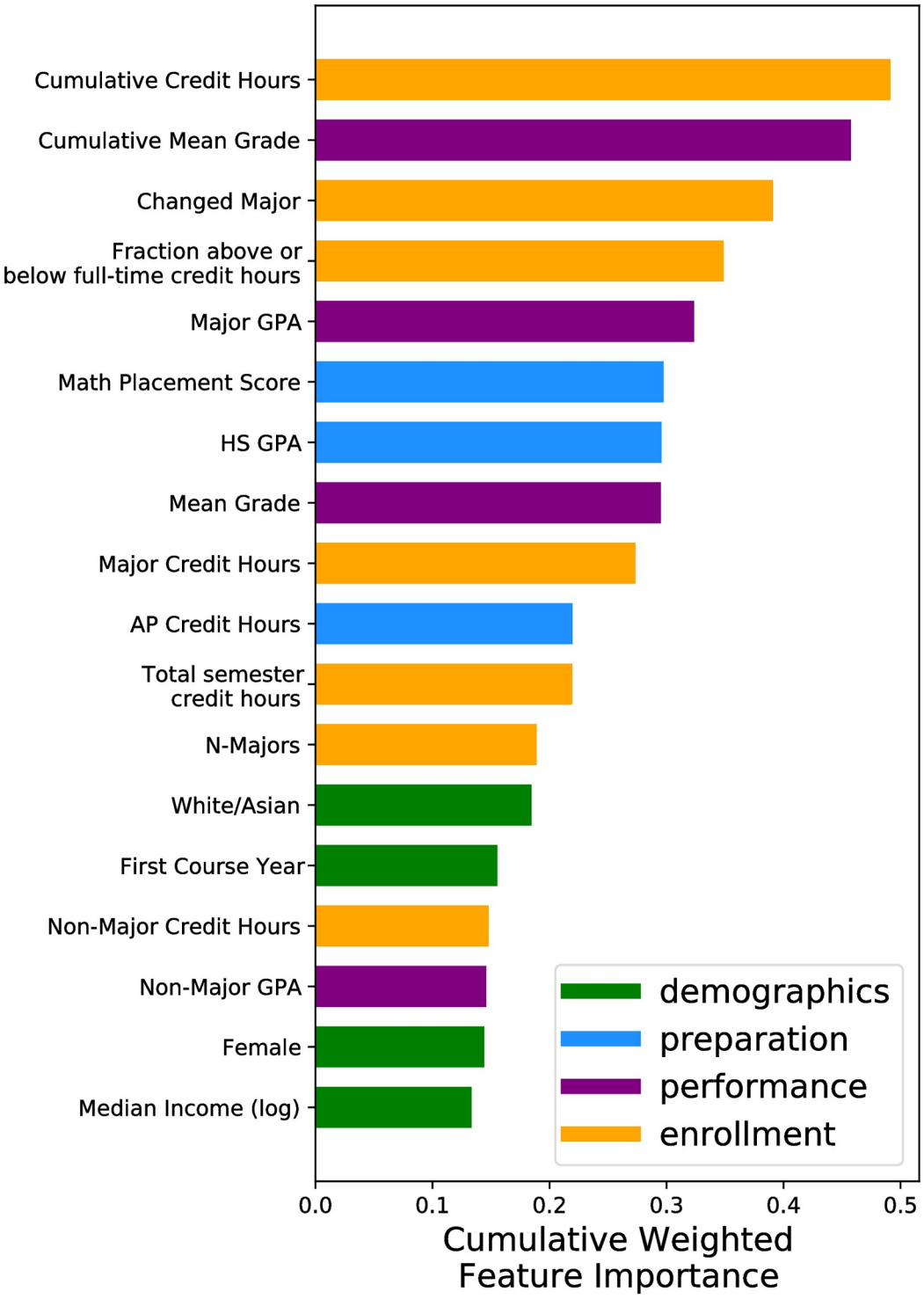

Fig 4. Average feature importance for all graduating semesters for the xgboost model. Feature importances are weighted by the F_1 scores per semester. Enrollment factors and the cumulative average grade are more likely to predict when a student graduates than other factors.

https://doi.org/10.1371/journal.pone.0242334.g004

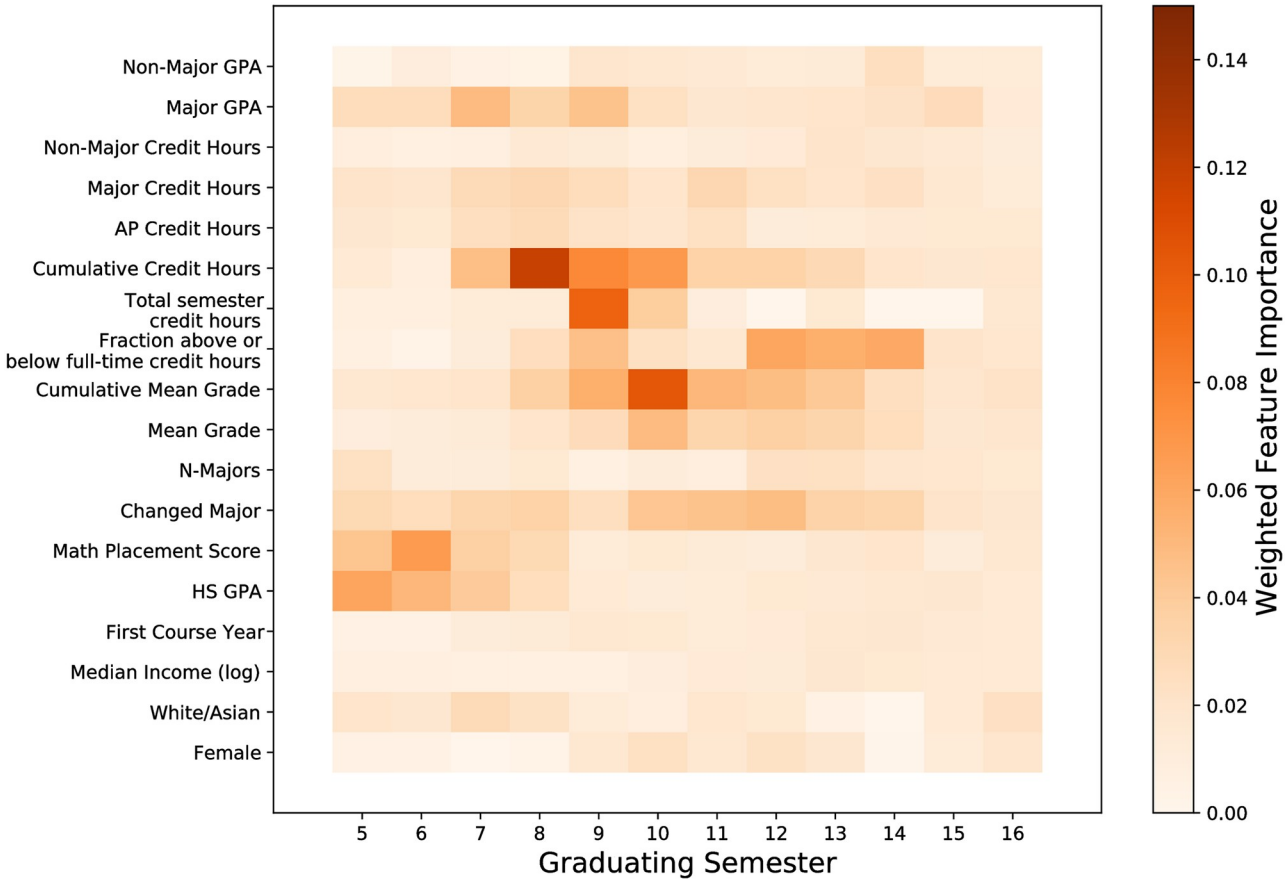

Fig 5. Xgboost feature importances for predicting if a student will graduate in the next semester. Feature importances are weighted by the F_1 score calculated from test data. The row-wise sum of these features would produce Fig 4. By far the most important feature for graduating "on time" (within 8 semesters) is having enough credit hours. Student preparation is important for graduating "early" (<8 semesters). Students who change majors are more likely to graduate later and thus this feature becomes more important in later graduating semesters.

https://doi.org/10.1371/journal.pone.0242334.g005

regression model (Fig 3). In the final semesters (semesters 15-16) logistic regression outperforms xgboost slightly.

Xgboost is more effective at correctly predicting the graduating semester of female students as assessed by the out-of-sample recall for all semesters except for semester 16 (Fig 9). Xgboost is also more effective at correctly predicting the graduating semester of under represented minority students for all semesters (Fig 9). Additionally, Xgboost is able to handle missing data cases better than the logistic regression model for all semesters (Fig 9). Due to the effectiveness of xgboost over logistic regression, the remaining results will focus on the xgboost model.

5.2 Effects on time-to-graduation

Tinto's Theory of Drop Out posits two overarching quantities as governing student choice to drop out or continue to graduation: 1) educational goal commitment, and 2) institutional commitment. Students initially start university with these commitments. These commitments are then mediated by a student's participating and integration into the academic system and the social system of a given institution. Generally speaking (Fig 10), graduating students take more credit hours per semester ($N(grad) = 15.1 \pm 0.01$, $N(no-grad) = 13.17 \pm 0.05$), enroll for

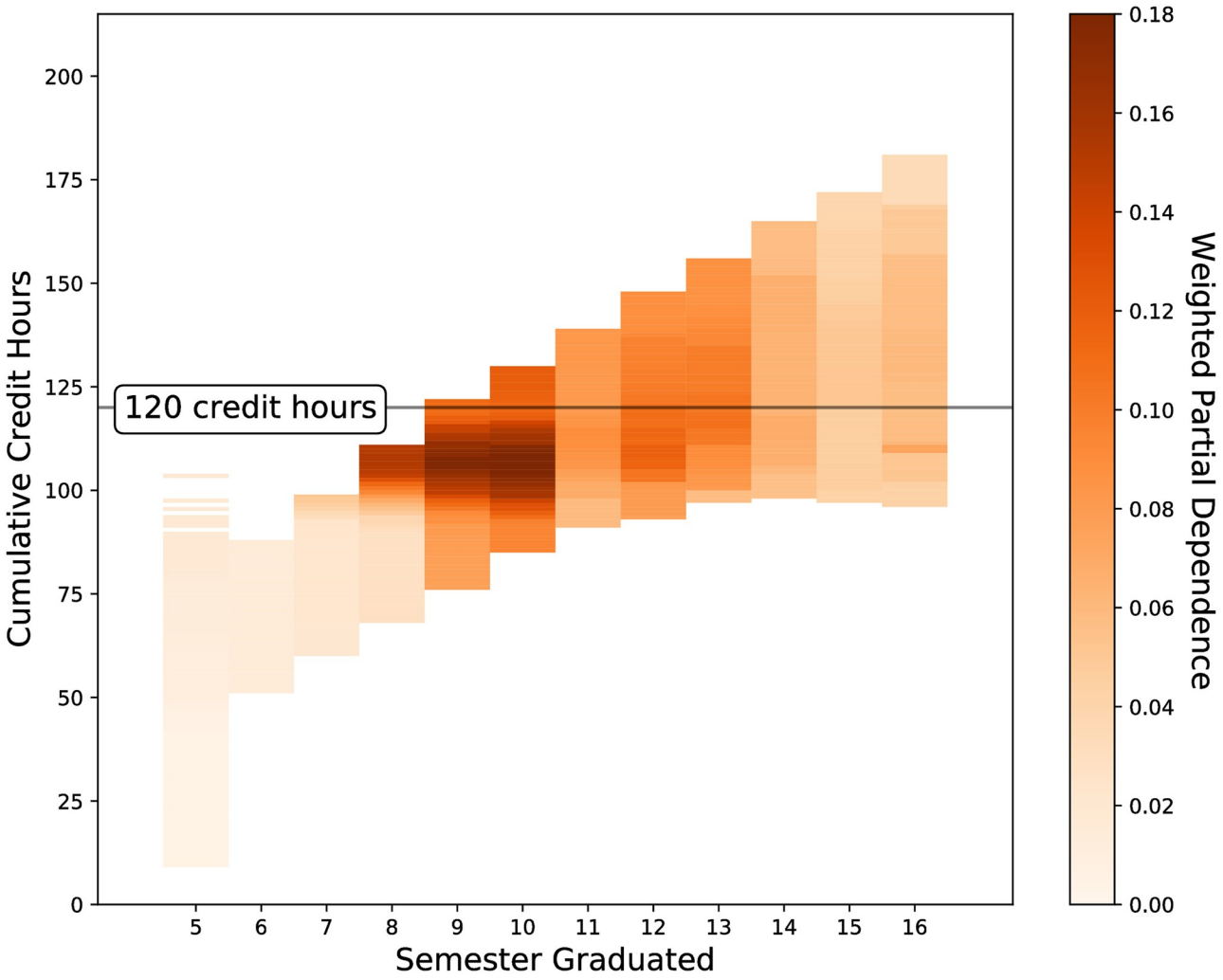

Fig 6. The partial dependence per semester for the cumulative credit hours a student has obtained. The stronger the partial dependence, the more contribution that value of cumulative credit hours has on the predicted probability that a student will graduate in the given semester. The partial dependence has been weighted by the per semester F_1 score. Having a total number of credit hours close to 120 is highly predictive of graduating if students graduate between 8 and 10 semesters of enrollment. Outside of this window the impact of the total number of credit hours diminishes. This is likely due to students having additional credit hours that do not count towards a degree such as when they change their major.

https://doi.org/10.1371/journal.pone.0242334.g006

more semesters ($N(grad) = 10.57 \pm 0.01$, $N(no - grad) = 8.79 \pm 0.05$), are more likely to be white or asian ($N(grad) = 86.7\%$, $N(no - grad) = 75.3\%$), and more likely to be female ($N(grad) = 54.2\%$, $N(no - grad) = 48.9\%$).

In this study, we characterize student's initial commitments as being mediated via the following quantities: high school GPA, math placement score, the number of AP credit hours the student possesses, the median income of the high school the student attended, gender, and race. Preparation is about half as important in predicting when a student graduates in comparison to enrollment factors (Fig 4). For early graduating students ($\leq$8 semesters), both math placement scores and high school GPAs were important in predicting that a student will graduate (Fig 5). While students are likely to have more AP credits on average if they graduate in 8-10 semesters ($AP_{8-10} = 4.06 \pm 0.03$, $AP_{other} = 2.26 \pm 0.01$), having AP credits was not an important indicator for predicting graduation during any semester in comparison to other features (Figs 4 and 5). While white and asian students and female students are more likely to graduate

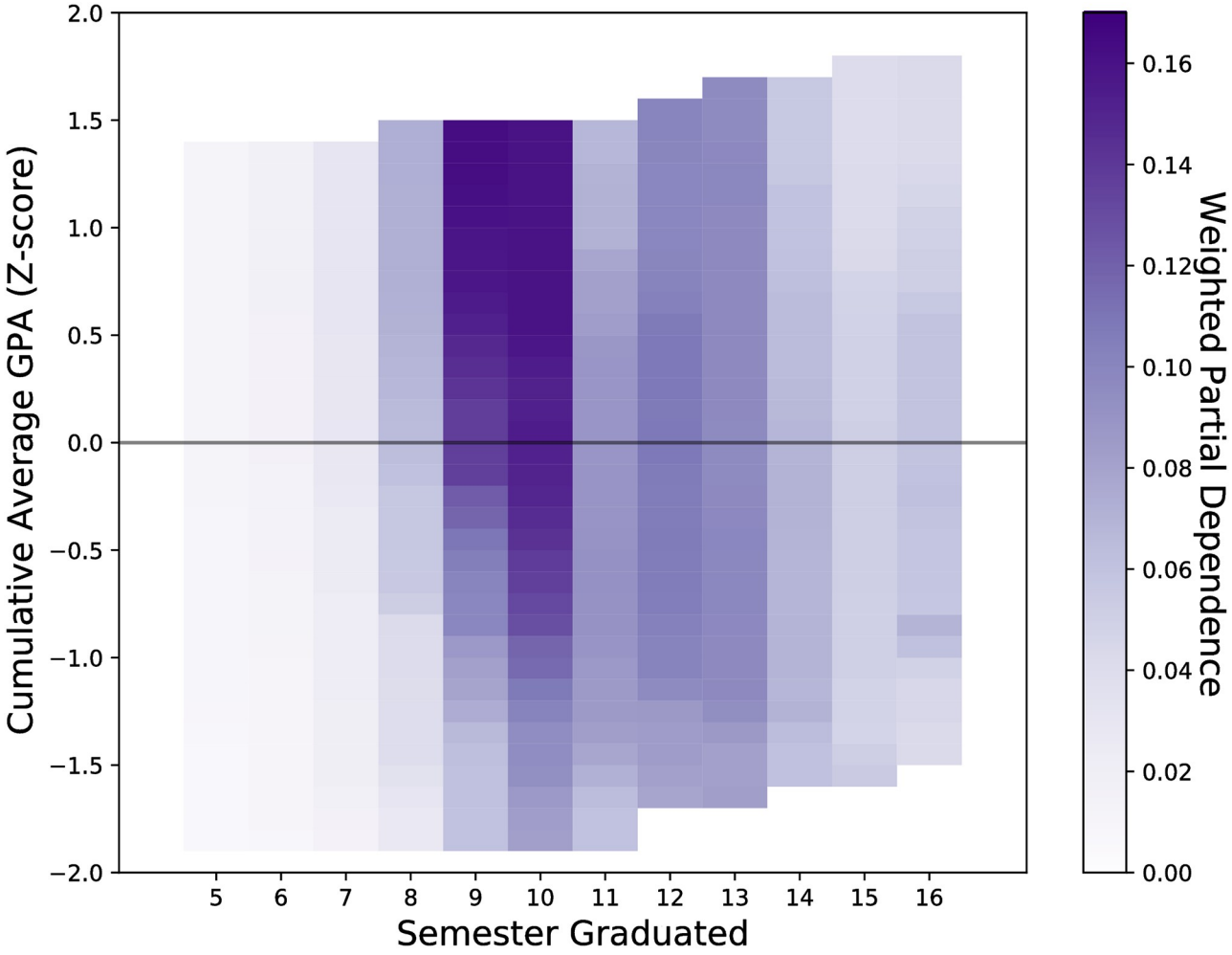

Fig 7. The partial dependence per semester for the cumulative average GPA a student has obtained. The stronger the partial dependence, the more contribution that the cumulative average GPA has on the predicted probability that a student will graduate in the given semester. The partial dependence has been weighted by the per semester F_1 score. Having a far above average GPA is a major contribution to graduating if a student graduates between semesters 9 and 10. This is likely due to these students never failing a course and having a high commitment to their chosen major.

https://doi.org/10.1371/journal.pone.0242334.g007

(Fig 10), demographics are not major indicators for predicting when a student graduates for any semester (Figs 4 and 5). This is also true for the median incomes of the zipcode the students went to high school.

In this study academic integration is represented by a combination of features including the cumulative credit hours, cumulative mean grade; and the per-semester fraction above or below full-time credit hour, mean grade, non-major GPA. A student's enrollment and performance are the most predictive features when predicting the semester when a student graduates (Figs 4 and 5). A student's cumulative credit hours is most important to predicting whether a student graduates "on-time" (after 8 semesters) or not and is overall important for prediction. The cumulative credit hours is less important for prediction the further a student's credits are away from 120 total credit hours (Fig 6). A student's average grade is most important for predicting that students will graduate within 10 semesters and is overall important for predicting when a student will graduate (Figs 4 and 5). Having above average grades is important for

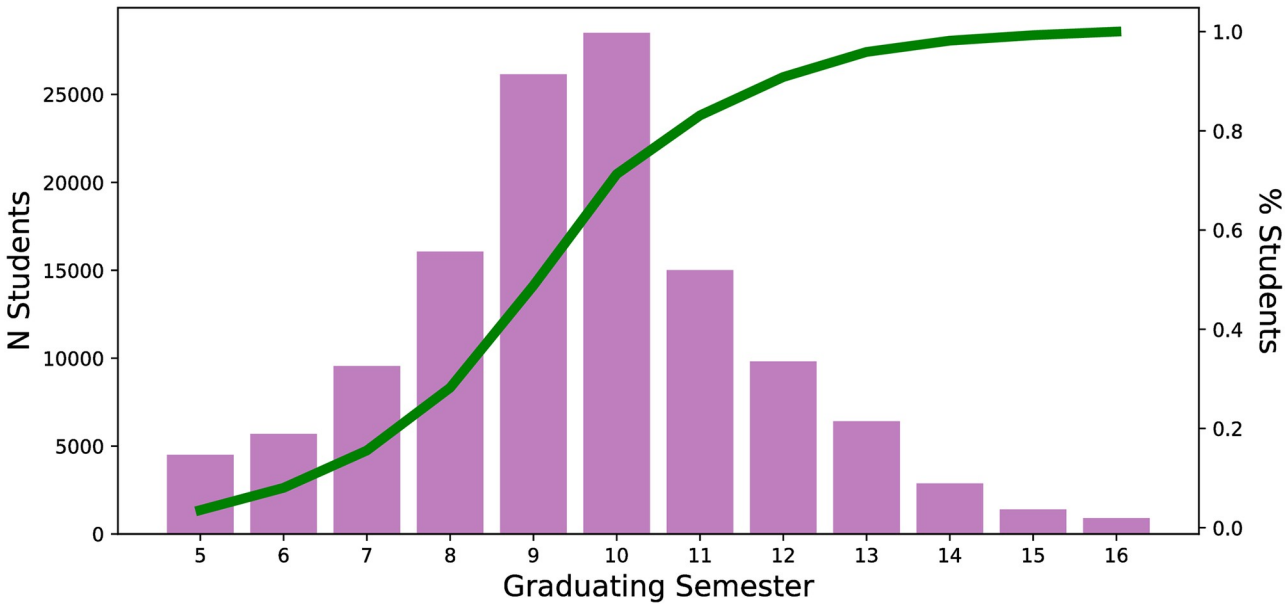

Fig 8. Students typically graduate within the window of 8 to 12 semesters. The bars represent the number of students across all cohort years who graduate in the following semester. The green line represents the cumulative fraction of students who have graduated.

https://doi.org/10.1371/journal.pone.0242334.g008

predicting students who graduate in 9-10 semesters however this is less important for other semesters (Fig 7).

In this study we use per semester major GPA, major credit hours, cumulative number of majors, and whether a student changed major in that semester or not to represent their social integration into the departments and learning communities associated with their chosen degree program. Student's performance within their major was not as an important indicator for predicting graduation in comparison to overall grade and cumulative credit hours (Fig 4). Changing a major has increased importance for predicting students who graduate later (Fig 5).

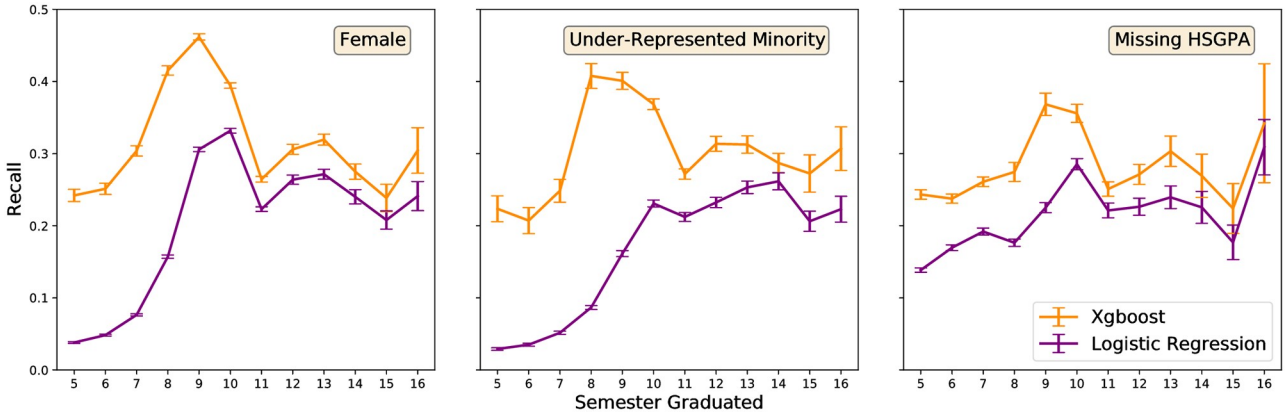

Fig 9. Recall scores for sub-populations within the data. The Xgboost model consistently labels women, under represented minorities, and students with missing data better than logistic regression. In Fig 3, there is a large disparity between the logistic regression model and the xgboost model for semesters 5-7. This may be due to the imputation method used with the logistic regression model and the correlation between graduating and missing data during those semesters. However, the gaps in model performance for women, minorities, and missing data in later semesters with no correlation are likely not due to imputation and demonstrate a strong preference for using the xgboost model over the logistic regression model.

https://doi.org/10.1371/journal.pone.0242334.g009

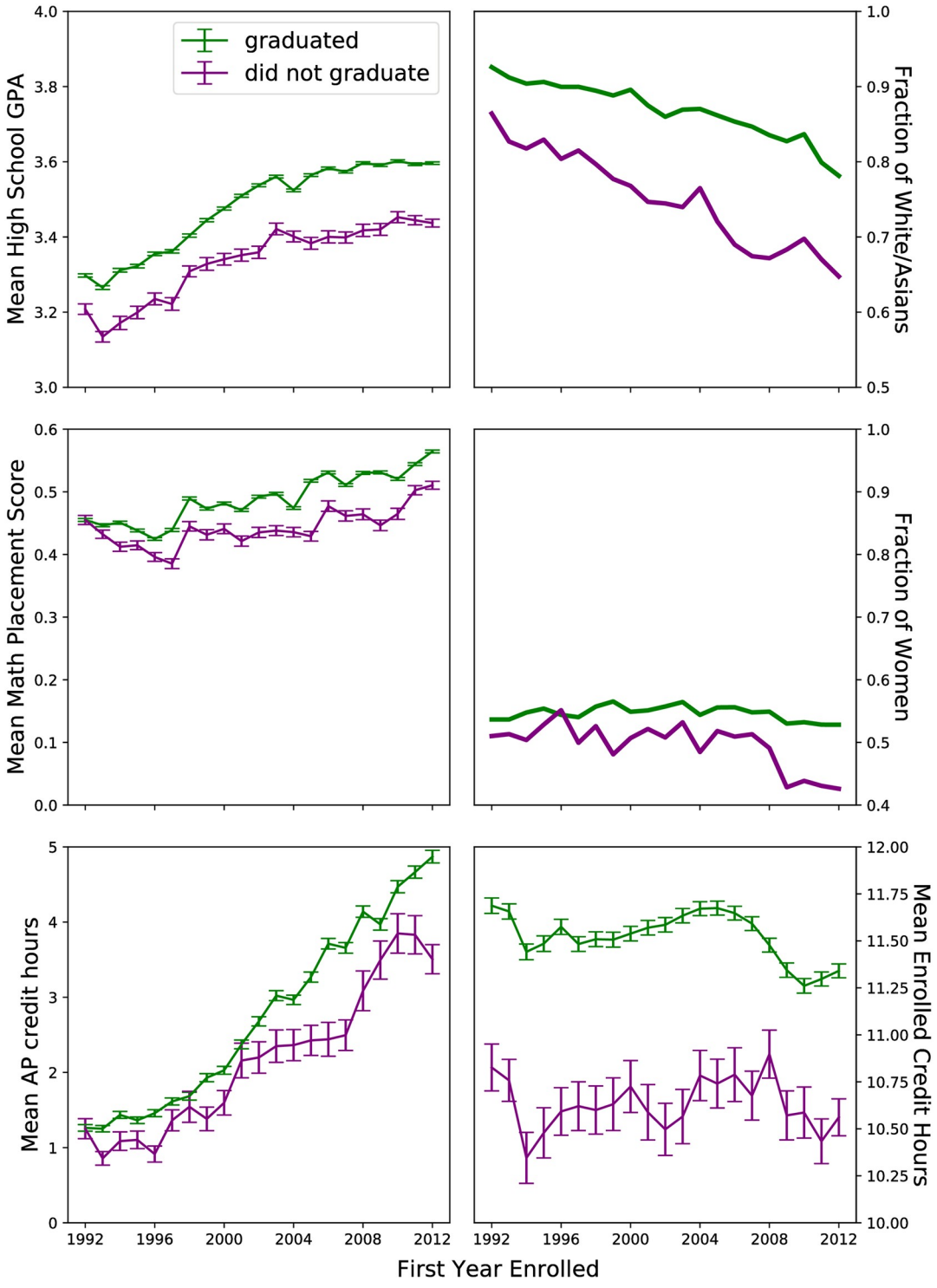

Fig 10. A comparison of a subset of features for students who graduate versus those who do not based on first year enrolled. Since 1992 the student population has become more diverse racially, students have had increasing high school GPAs and math placement scores, and the time it takes students who graduate to graduate has been decreasing. Students who graduate typically take more credit hours than those who don't, typically are better prepared as measured by high school GPA and math placement score, and are less diverse than the total university population.

https://doi.org/10.1371/journal.pone.0242334.g010

5.3 Gradient boosting

The models in this study attempt to predict whether a student will graduate or not during the observation period of being enrolled for 5-16 semesters. The xgboost model is generally more effective than the logistic regression model except in the last semesters (15-16) studied as assessed by the out-of-sample F_1-score (Fig 3). Xgboost is particularly more effective in highly imbalanced case of students graduating (approximately 5% per semester (Fig 8)) within $\leq$8 semesters (Fig 3). In the more balanced case (semesters 8-14 have approximately 20% of the students eligible to graduate do so (Fig 8)), Xgboost still performs better than the logistic regression model (Fig 3). In the final semesters (semesters 15-16) logistic regression outperforms xgboost slightly.

Xgboost is more effective at correctly predicting the graduating semester of female students as assessed by the out-of-sample recall for all semesters except for semester 16 (Fig 9). Xgboost is also more effective at correctly predicting the graduating semester of under represented minority students for all semesters (Fig 9). Additionally, Xgboost is able to handle missing data cases better than the logistic regression model for all semesters (Fig 9). Due to the effectiveness of xgboost over logistic regression, the remaining results will focus on the xgboost model.

6 Discussion

This paper presents two complimentary results: 1) gradient boosting is a useful tool for predicting when students graduate in comparison to traditional statistical algorithms, 2) students who actively integrate into their academic and social communities is the primary effect that predicts when a student graduates thus following Tinto's Theory. 1) students who actively integrate into their academic and social communities is the primary effect that predicts when a student graduates thus following Tinto's Theory, 2) gradient boosting is a useful tool for predicting when students graduate in comparison to traditional statistical algorithms. This section will discuss why the xgboost model outperforms logistic regression, the implications that come with the results of the model, and compare it to other studies that have predicted time-to-graduation using Tinto's Theory of Drop Out and hazard modeling more broadly.

6.1 Why does gradient boosting produce better fitted models than maximum likelihood estimation?

In almost every case, the xgboost model predicts graduation better than the logistic regression model. This can be for several reasons. First, the xgboost model fitting procedure is a slower iterative procedure (1000 iterations) than the maximum likelihood estimation (typically 5-6 iterations) that the logistic regression model uses. Thus, in each iteration the xgboost model can focus on the local neighborhood of feature details within the training data that the logistic regression model may miss [61].

Second, the xgboost model uses a custom imputation engine [10]. Whenever data is missing and an learner is using a feature with missing data, the missing data is imputed to be the most likely choice in the decision tree. Because there are many learners, they can account for different local patterns in the data. The data for the logistic regression model is imputed from the mean of the student's starting year cohort. This mean imputation likely loses information that the xgboost model is able to attend to. It could be that a more sophisticated imputation model will increase the logistic regression model performance. Typically if data missingness correlates with the outcome variable, then data should be imputed to prevent bias due to the missing data [62]. In our case, data missingness for both high school GPA and math placement score

correlates with early graduation rates (see S1 Fig). Thus the logistic regression model performance would likely increase for semester's 5-7 with a more sophisticated imputation model. However this would not explain the substantial improvements xgboost makes over the logistic model for women and under represented minorities in later semesters.

Third, xgboost penalizes leafs within the tree learners that are fit on few examples from the training data [10]. Additionally, xgboost weights on class labels as well. This is especially useful in the highly unbalanced case of predicting when a student graduates. Because the weight of training data from students who do not graduate is tuned via a grid search, this prevents the xgboost model from overfitting on the majority class simply due to having more representative samples.

6.2 Effects on time-to-graduation

Tinto's theory suggests that students have an initial level of intent to graduate from an institution upon entry [2, 4]. This intent is the combination of a students family background (e.g., socio-economic status), individual attributes (e.g., academic ability, race), and pre-college experiences (e.g., high school GPA). This intent is then tempered by the student's at-college experience such as social integration [63], financial support [9], and academic performance.

6.2.1 Effects on the initial conditions of educational commitment. Student's backgrounds can set up a wide variety of contributions to their initial educational and institutional commitments. In this study we assess these initial conditions using a student's high school GPA, math placement score, the number of AP credit hours the student possesses, the median income of the high school the student attended, gender, and race. This university uses a math placement test to determine a student's incoming math ability to place them in the appropriate math course. This test was designed at the university and has been used for the entire study period. This test was not designed with psychometrics in mind. Thus, while the test is representative of some measurement of math ability, it is unclear how much the test is representative of math ability. However, we see that a student's math placement score is the most important attribute that predicts when a student will graduate that is not a performance or enrollment variable Fig 5). This could be similar to [13] evidence that the starting math course is very important to staying in STEM. In [13], students taking lower level math courses at college had increased likelihood of leaving the STEM major they were enrolled in for a different major. In our case, students who take remedial math courses simply have more courses to take and thus must remain at the university longer. Given that students in our study who come from higher income communities are able to remain at university longer before graduating, this indicates that remedial mathematics should be examined more in terms of the cost for a student. It should also be noted that in this study we are predicting all students time to graduation even those who are not pursuing quantitative degrees. The result that math ability has a dominant effect in the context of other demographic and preparation variables is consistent with the literature which claims that math ability has an sizeable effect on student performance at university [16].

A student's high school GPA is similarly important to a student's math ability in predicting when they graduate (Figs 4 and 5). This is consistent with some literature that finds that students with higher high school GPAs are more likely to graduate within 4 years [9]. [8] found in a similar study that high school GPA had little to no effect on predicting when a student drops out from university. In both this paper and [8] the study uses data from university's that serve primarily students who come from the state the university resides in. Thus this may indicate that the effect of high school GPA on college success is geographically dependent on the quality of high school preparation for university.

Perhaps counter-intuitively having AP credits is not as important as a student's high school GPA or math ability (Fig 4). AP credits directly count for college credits. Students with more AP credit have fewer credit hours necessary to graduate. In this study, having AP credits is an indicator that a student will graduate in four years (8-9 semesters). However it is not a strong indicator that a student will graduate early in comparison to a student's math placement score and high school GPA. [8] found that having transfer credits had no effect on on students dropping out. This was also true in a study of physics students who change their majors [21]. It could be that AP credits are too random in whether they count for credit or not. A student may choose to take a course anyways they have AP credit for if its in a sequence (e.g., introductory physics) because they may feel ill prepared for the second semester course. It could also be that some students who take AP courses do so with the intent to take a minor or dual major. In this case, the AP courses "free" up more time at university to be able to graduate in four years.

Financial aid is one of the top reasons students leave the university without a degree [9]. In this study we do not know if a student has access to financial aid or not. However, we do include the median income for families that live in the zip code of the high school the student attended according to the 2011 census American Community Survey 5- year estimate [47]. This is a rough estimate of the socio-economic status of the high school that the student attends. A course grained measurement like this does not capture all of the nuances of individual students financial support and in our case, we see that this feature was less important in comparison to performance and enrollment features. In our study, we find that coming from higher income high school's is an indicator that a student will take longer to graduate. [48] found that the socio-economic status of a student's peers had an approximately equal effect as a student's individual socio-economic status on high school student exit examination scores. While this effect is small, it may indicate that students from higher income regions have access to more resources and thus can spend more time in college before entering the work force. This is a similar result to [9] which found that students who come from low income households are likely to graduate within 4 years in comparison to students from higher income households.

6.2.2 Dynamic contributions to a student's academic integration. A student's academic integration is defined by Tinto [4] as being a combination of performance in university and their intellectual development. In this study we characterize performance through both cumulative GPA measurements and per semester GPA measurements. Further, we split this into major and non-major GPAs.

Performance as measured by GPA has had a demonstrated primary effect on graduating [1, 8, 14]. [8] found that there was a decaying impact on drop out due to GPA. The longer a student was enrolled, the less their GPA was likely to be a strongly influencing factor on dropping out. In our study we find somewhat different results, namely students with above average cumulative GPAs are more likely to graduate on time (Fig 7). However this effect diminishes with time. It could be that this peak effect is due to two reasons: 1) high performing students are more likely to graduate on time [1], and 2) failing a single course significantly sets back both the time to graduate and a student's cumulative GPA.

There is an interplay between the cumulative credit hours and the cumulative mean grade. In semester 8, the single most predictive feature is cumulative credit hours (Fig 5). However by semester 10, cumulative credit hours has exchanged the highest rank with the cumulative mean grade. This has a couple implications. First, the decaying impact of grades as noted by [8] in this case begins later at semester 10. Second, it may be that to graduate within 8 semesters (4 years), there are very few paths other than a strictly laid out course schedule with no

deviations and no failing grades. Whereas graduating within 5 years allows more leeway for students to take additional courses that could be due to, for example, receiving a minor.

Second, there is a transition across graduating semesters of what is important (Fig 5). Early graduation is predicated more by a student's initial conditions than anything else. By semester 8, the overwhelming effect on successfully predicting students graduating in this semester is their cumulative credit hours. Past semester 8, there is a strong combination of features that are predictive of graduating. This suggests that while Tinto's theory indicates that there is a dynamic contribution of on-campus interactions to student educational and institutional commitments, these dynamic contributions may not matter that much for students with very strong backgrounds who wish to graduate early.

6.2.3 Dynamic contributions to a student's social integration. A student's social integration is defined by Tinto [4] as being a combination of peer group interactions and faculty interactions. In this study we use a course grained measurement of peer group and faculty interactions by measuring the total per semester credit hours a student registers for in their major. Additionally, we note when a student changes their major. Changing a major is a distinct situation where a student will leave their peer and faculty group for a new group and thus may indicate low social integration.

In this study students who attend the university and take a recommended load of courses and pass each course were likely to graduate in 8-10 semesters. When students altered from this path (e.g., a student changes their major) they increased the amount of time it took to graduate or did not graduate during the study period. [13] provides, at least for STEM students, a complementary explanation as to why a student may graduate later. This university has a very large enrollment (typically $>$50000 students), a strong research program, and a strong greek life. In each case these may contribute to social integrations given there is more opportunity at this university than some others for meeting new people, participating in research, or participating in social events.

While not being the largest effect, students who take higher amounts of major credit hours were more likely to graduate in semesters 8 and 9 (Fig 5). This is sensible given's Tinto's theory since taking more major credit hours both works toward's graduation and is an indication of a strong integration into a peer and faculty group. Within the literature, STEM students who took fewer STEM courses in the first year, were enrolled in less challenging math courses in the first year, and performed poorly in their STEM courses in comparison to non-STEM courses were highly likely to switch majors. Tinto would describe these students as having a reduced educational commitment due to not integrating into the social community and/or lack in the academic engagement of their chosen degree program. In our study we find similar conclusions, higher performance and the number of major credit hours is a more likely indicator of graduating in four years and this effect diminishes over time after the 4 year (8-9 semesters) mark (Fig 5). Thus, it may be likely that performance in a major and the frequency of major courses taken may be an indicator of lower social and academic engagement described by Tinto [4].

In this study we have highlighted that changing a major can have a profound effect on the time it takes to graduate. Changing a major impacts a student's time-to-graduation and is predictive of students who graduate later (Fig 5). A student who changes their major could do so due to low academic integration [13]. However this low academic integration could represent low social integration as well [23]. Student's who perform poorly may ask themselves if they "belong" in a major. In many cases STEM students who demonstrate high academic performance change their major due to reasons associated with social and personal interactions at university [13, 23]. This especially affects women and under represented minorities [23]. In this study changing a major does not show a strong correlation with a student's race or gender

($\rho_{race} = -0.07$, $\rho_{gender} = -0.03$). In many cases, the experiences that lead to changing a major are more nuanced than a single variable can contain thus while race and gender are a strong component to major change as reported in qualitative interviews [23], this may not be captured in binary variables. Ultimately, changing majors has been a poorly investigated and deserves more investigation. Future work will look at using Tinto's theory as a framework for investigating student's changing their major.

6.3 Why does gradient boosting produce better fitted models than maximum likelihood estimation?

In almost every case, the xgboost model predicts graduation better than the logistic regression model. This can be for several reasons. First, the xgboost model fitting procedure is a slower iterative procedure (1000 iterations) than the maximum likelihood estimation (typically 5-6 iterations) that the logistic regression model uses. Thus, in each iteration the xgboost model can focus on the local neighborhood of feature details within the training data that the logistic regression model may miss [61].

Second, the xgboost model uses a custom imputation engine [10]. Whenever data is missing and an learner is using a feature with missing data, the missing data is imputed to be the most likely choice in the decision tree. Because there are many learners, they can account for different local patterns in the data. The data for the logistic regression model is imputed from the mean of the student's starting year cohort. This mean imputation likely loses information that the xgboost model is able to attend to. It could be that a more sophisticated imputation model will increase the logistic regression model performance. Typically if data missingness correlates with the outcome variable, then data should be imputed to prevent bias due to the missing data [62]. In our case, data missingness for both high school GPA and math placement score correlates with early graduation rates (see S1 Fig). Thus the logistic regression model performance would likely increase for semester's 5-7 with a more sophisticated imputation model. However this would not explain the substantial improvements xgboost makes over the logistic model for women and under represented minorities in later semesters.

Third, xgboost penalizes leafs within the tree learners that are fit on few examples from the training data [10]. Additionally, xgboost weights on class labels as well. This is especially useful in the highly unbalanced case of predicting when a student graduates. Because the weight of training data from students who do not graduate is tuned via a grid search, this prevents the xgboost model from overfitting on the majority class simply due to having more representative samples.

6.4 Limitations on this study

Given that a perfect F_1-score is 1 and in this paper we report F_1-scores around 0.5 at the maximum, it is likely that some of these effects such as financial aid or social integration have large effects on when a student graduates. In this paper we have no direct measurement of social integration. We have used proxy measurements, such as the credit hours a student has accumulated towards their currently registered major program, as data towards the student's social integration. Ideally this would have increased fittedness in the models as these variables may represent some assessment of social integration. This proved not to be the case. The models still have middling fit values which is likely due to this study having no access to direct measurements of social integration. Social interaction data is difficult to come by without actively observing students through classroom observation, surveys, etc. Future work will leverage new methods created for using passively collected data to assess student social network status as a direct measurement of social integration [64]. Additionally, it is common in some studies,

(e.g., [1]) to use individual heterogeneity models or "frailty" models to assess unmeasured contributions to graduation [65]. In our case, we attempted initially to use gamma distributed frailty terms [11, 66]. These proved to provide no increase in model performance thus were removed from subsequent analysis. Future work will consider what contributes to the unobserved heterogeneity (e.g., direct measurements of peer social engagement such as social network centrality).

Additionally, this paper has focused on a student's educational commitment (per [4]) as opposed to their institutional commitment. Much of this is mediated by the fact that 88% of students in the study do in fact graduate, there may be a subset of students who leave simply due to the fact they become disillusioned with the institution and want to pursue a degree elsewhere. Thus there is likely other factors that are not observed that effects when a student a graduates. Many of these factors, such as life events such as family hardship, are necessarily hard to observe for an entire university population. There are also likely factors that are particular to this university that are not common or not impacting at other institutions. Future work will investigate the effect of student social network belonging and the amount of financial aid available to the students.

7 Conclusion

This paper has presented a discrete time hazard model predicting when a student will graduate. It uses a novel method to calculate the logistic regression called gradient boosting which has been shown to provide better fits than traditional maximum likelihood. While using more sophisticated imputation with the logistic regression model might have increased performance, the ease of use of xgboosts built-in imputation engine provides a strong advantage in it's use to researchers. Given the utility of gradient boosting in this setting, especially in providing better predictions for under-served populations, this paper recommends that this method be more prevalent in the education research community. Additionally, other methods should be examined such as artificial neural networks [67].

Additionally, this paper used the partial dependence method to examine two of the model variables. Partial dependence tells us the contributions to the predicted probability of the model for the entire feature space. This method allows us to examine the entirety of continuous variables such as GPA instead of reducing them to a single value associated with a model coefficient. This method too should see much broader use in the education research community.

This paper follows Tinto's theory of drop out that predicts social integration and college participation are more likely to impact a student's commitment to graduation than second order effects such as preparation. Future work will connect student academic performance, participation in course work, with social network metrics and financial aid information. Future work will also examine the specific effects that remedial math have on student retention in a major, their time to graduation, and if participating in remedial math courses lowers the likelihood to graduate.

Supporting information

S1 Fig. The fraction of missing data and the spearman rank correlation between whether data is missing and the outcome variable of graduating the following semester. There is a small correlation in the early semesters (5-7) between whether students are graduating and if they have missing data or not.
(TIF)

Author Contributions

Conceptualization: John M. Aiken, Riccardo De Bin.

Data curation: John M. Aiken, Marcos D. Caballero.

Formal analysis: John M. Aiken, Riccardo De Bin.

Funding acquisition: Morten Hjorth-Jensen, Marcos D. Caballero.

Investigation: John M. Aiken.

Methodology: John M. Aiken, Riccardo De Bin.

Project administration: John M. Aiken.

Resources: John M. Aiken.

Software: John M. Aiken.

Supervision: Morten Hjorth-Jensen, Marcos D. Caballero.

Validation: John M. Aiken.

Visualization: John M. Aiken.

Writing – original draft: John M. Aiken, Riccardo De Bin.

Writing – review & editing: John M. Aiken, Riccardo De Bin, Morten Hjorth-Jensen, Marcos D. Caballero.

References

1. Yue Hongtao and Fu Xuanning. "Rethinking graduation and time to degree: A fresh perspective". In: *Research in Higher Education* 58.2 (2017), pp. 184–213. https://doi.org/10.1007/s11162-016-9420-4
2. Braxton John M, Milem Jeffrey F, and Sullivan Anna Shaw. "The influence of active learning on the college student departure process: Toward a revision of Tinto's theory". In: *The journal of higher education* 71.5 (2000), pp. 569–590. https://doi.org/10.2307/2649260
3. Abel Jaison R, Deitz Richard, and Su Yaqin. "Are recent college graduates finding good jobs?" In: *Current issues in economics and finance* 20.1 (2014).
4. Tinto Vincent. "Dropout from higher education: A theoretical synthesis of recent research". In: *Review of educational research* 45.1 (1975), pp. 89–125. https://doi.org/10.3102/00346543045001089
5. Pascarella Ernest T and Terenzini Patrick T. "Predicting voluntary freshman year persistence/withdrawal behavior in a residential university: A path analytic validation of Tinto's model." In: *Journal of educational psychology* 75.2 (1983), p. 215. https://doi.org/10.1037/0022-0663.75.2.215
6. Nora Amaury, Attinasi LC Jr, and Matonak Andrew. "Testing qualitative indicators of precollege factors in Tinto's attrition model: A community college student population". In: *The Review of Higher Education* 13.3 (1990), pp. 337–355. https://doi.org/10.1353/rhe.1990.0021
7. Cabrera Alberto F, Nora Amaury, and Castaneda Maria B. "College persistence: Structural equations modeling test of an integrated model of student retention". In: *The journal of higher education* 64.2 (1993), pp. 123–139. https://doi.org/10.1080/00221546.1993.11778419
8. DesJardins Stephen L, Ahlburg Dennis A, and McCall Brian P. "An event history model of student departure". In: *Economics of education review* 18.3 (1999), pp. 375–390. https://doi.org/10.1016/S0272-7757(98)00049-1
9. Ishitani Terry T. "A longitudinal approach to assessing attrition behavior among first-generation students: Time-varying effects of pre-college characteristics". In: *Research in higher education* 44.4 (2003), pp. 433–449. https://doi.org/10.1023/A:1024284932709
10. Tianqi Chen and Carlos Guestrin. "Xgboost: A scalable tree boosting system". In: *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*. ACM. 2016, pp. 785–794.
11. Yamaguchi Kazuo. *Event history analysis*. Vol. 28. Sage, 1991.

12. Scott Marc A and Kennedy Benjamin B. "Pitfalls in pathways: Some perspectives on competing risks event history analysis in education research". In: *Journal of Educational and Behavioral Statistics* 30.4 (2005), pp. 413–442. https://doi.org/10.3102/10769986030004413

13. Xianglei Chen. *STEM Attrition: College Students' Paths into and out of STEM Fields. Statistical Analysis Report. NCES 2014-001*. Tech. rep. 2013. URL: https://nces.ed.gov/pubs2014/2014001rev.pdf.

14. Chen Rong. "Institutional characteristics and college student dropout risks: A multilevel event history analysis". In: *Research in Higher education* 53.5 (2012), pp. 487–505. https://doi.org/10.1007/s11162-011-9241-4

15. Trusty Jerry and Niles Spencer G. "High-school math courses and completion of the bachelor's degree". In: *Professional School Counseling* (2003), pp. 99–107.

16. Gaertner Matthew N et al. "Preparing students for college and careers: The causal role of algebra II". In: *Research in Higher Education* 55.2 (2014), pp. 143–165. https://doi.org/10.1007/s11162-013-9322-7

17. Bamberg Betty. "Composition instruction does make a difference: A comparison of the high school preparation of college freshmen in regular and remedial English classes". In: *Research in the Teaching of English* 12.1 (1978), pp. 47–59.

18. Sadler Philip M and Tai Robert H. "Success in introductory college physics: The role of high school preparation". In: *Sci. Educ.* 85.2 (2001), pp. 111–136. https://doi.org/10.1002/1098-237X(200103)85:2%3C111::AID-SCE20%3E3.0.CO;2-O

19. Hazari Zahra, Tai Robert H, and Sadler Philip M. "Gender differences in introductory university physics performance: The influence of high school physics preparation and affective factors". In: *Science Education* 91.6 (2007), pp. 847–876. https://doi.org/10.1002/sce.20223

20. Zwick Rebecca and Sklar Jeffrey C. "Predicting college grades and degree completion using high school grades and SAT scores: The role of student ethnicity and first language". In: *American Educational Research Journal* 42.3 (2005), pp. 439–464. https://doi.org/10.3102/00028312042003439

21. Aiken John M, Henderson Rachel, and Caballero Marcos D. "Modeling student pathways in a physics bachelor's degree program". In: *Physical Review Physics Education Research* 15.1 (2019), p. 010128. https://doi.org/10.1103/PhysRevPhysEducRes.15.010128

22. Zabriskie Cabot et al. "Using machine learning to predict physics course outcomes". In: *Physical Review Physics Education Research* 15.2 (2019), p. 020120. https://doi.org/10.1103/PhysRevPhysEducRes.15.020120

23. Seymour Elaine. *Talking about leaving: Why undergraduates leave the sciences*. Westview Press, 2000.

24. Alexander Karl L et al. "Social background, academic resources, and college graduation: Recent evidence from the National Longitudinal Survey". In: *American Journal of Education* 90.4 (1982), pp. 315–333. https://doi.org/10.1086/443651

25. Sewell William H and Shah Vimal P. "Socioeconomic status, intelligence, and the attainment of higher education". In: *Sociology of education* (1967), pp. 1–23. https://doi.org/10.2307/2112184

26. Smith William Richard, Edmister Julie, and Sullivan Kathleen. "Factors influencing graduation rates at Mississippi's public universities". In: *College and University* 76.3 (2001), p. 11.

27. Rowan-Kenyon Heather T. "Predictors of delayed college enrollment and the impact of socioeconomic status". In: *The Journal of Higher Education* 78.2 (2007), pp. 188–214. https://doi.org/10.1353/jhe.2007.0012

28. Susan P Choy. *Students whose parents did not go to college: Postsecondary access, persistence, and attainment*. 2001.

29. Pascarella Ernest T et al. "First-generation college students: Additional evidence on college experiences and outcomes". In: *The Journal of Higher Education* 75.3 (2004), pp. 249–284. https://doi.org/10.1353/jhe.2004.0016

30. Hastie Trevor et al. *The elements of statistical learning: data mining, inference and prediction*. Vol. 27. 2. Springer, 2005, pp. 83–85.

31. Skipper Seabold and Josef Perktold. "statsmodels: Econometric and statistical modeling with python". In: *9th Python in Science Conference*. 2010.

32. R Core Team. *R: A Language and Environment for Statistical Computing*. ISBN 3-900051-07-0. R Foundation for Statistical Computing. Vienna, Austria, 2020. URL: http://www.R-project.org/.

33. Biihlmann Peter, Hothorn Torsten, et al. "Boosting algorithms: Regularization, prediction and model fitting". In: *Statistical Science* 22.4 (2007), pp. 477–505. https://doi.org/10.1214/07-STS242

34. Mayr Andreas et al. "Extending statistical boosting". In: *Methods of information in medicine* 53.06 (2014), pp. 428–435. https://doi.org/10.3414/ME13-01-0123

35. Biihlmann Peter and Yu Bin. "Boosting with the L 2 loss: regression and classification". In: *Journal of the American Statistical Association* 98.462 (2003), pp. 324–339. https://doi.org/10.1198/016214503000125

36. Yuk Lai Suen, Prem Melville, and Raymond J Mooney. "Combining bias and variance reduction techniques for regression trees". In: *European Conference on Machine Learning*. Springer. 2005, pp. 741–749.

37. Lombardo L et al. "Binary logistic regression versus stochastic gradient boosted decision trees in assessing landslide susceptibility for multiple-occurring landslide events: application to the 2009 storm event in Messina (Sicily, southern Italy)". In: *Natural Hazards* 79.3 (2015), pp. 1621–1648. https://doi.org/10.1007/s11069-015-1915-3

38. Xu Qian et al. "PDC-SGB: Prediction of effective drug combinations using a stochastic gradient boosting algorithm". In: *Journal of theoretical biology* 417 (2017), pp. 1–7. https://doi.org/10.1016/j.jtbi.2017.08.019 PMID: 28099868

39. Ma Xiaolei et al. "Prioritizing influential factors for freeway incident clearance time prediction using the gradient boosting decision trees method". In: *IEEE Transactions on Intelligent Transportation Systems* 18.9 (2017), pp. 2303–2310. https://doi.org/10.1109/TITS.2016.2635719

40. De Menezes Fortunato S et al. "Data classification with binary response through the Boosting algorithm and logistic regression". In: *Expert Systems with Applications* 69 (2017), pp. 62–73. https://doi.org/10.1016/j.eswa.2016.08.014

41. Bryk Anthony S and Raudenbush Stephen W. "Toward a more appropriate conceptualization of research on school effects: A three-level hierarchical linear model". In: *Multilevel analysis of educational data*. Elsevier, 1989, pp. 159–204.

42. Friedman Jerome H. "Stochastic gradient boosting". In: *Computational statistics & data analysis* 38.4 (2002), pp. 367–378. https://doi.org/10.1016/S0167-9473(01)00065-2

43. Pascarella Ernest T, Terenzini Patrick T, and Wolfle Lee M. "Orientation to college and freshman year persistence/withdrawal decisions". In: *The Journal of Higher Education* 57.2 (1986), pp. 155–175. https://doi.org/10.2307/1981479

44. M Scott DeBerard, Spielmans Glen I, and Julka Deana L. "Predictors of academic achievement and retention among college freshmen: A longitudinal study". In: *College student journal* 38.1 (2004), pp. 66–81.

45. *IPEDs definitions*. https://nces.ed.gov/ipeds/report-your-data/race-ethnicity-reporting-changes. Accessed: 10-23-2018.

46. Poldrack Russell A, Huckins Grace, and Varoquaux Gael. "Establishment of Best Practices for Evidence for Prediction: A Review". In: *JAMA psychiatry* (2019).

47. *American Community Survey Tables: 2007—2011 B19013A*. URL: https://factfinder.census.gov/bkmk/table/1.0/en/ACS/11_5YR/S1903/0100000US.86000.

48. Caldas Stephen J and Bankston Carl. "Effect of school population socioeconomic status on individual academic achievement". In: *The Journal of Educational Research* 90.5 (1997), pp. 269–277. https://doi.org/10.1080/00220671.1997.10544583

49. Friedman Jerome H. "Greedy function approximation: a gradient boosting machine". In: *Annals of statistics* (2001), pp. 1189–1232. https://doi.org/10.1214/aos/1013203451

50. Aurelien Geron. *Hands-on machine learning with Scikit-Learn and TensorFlow: concepts, tools, and techniques to build intelligent systems*. "O'Reilly Media, Inc.", 2017. URL: https://dl.acm.org/citation.cfm?id=3153997.

51. Bergstra James and Bengio Yoshua. "Random search for hyper-parameter optimization". In: *Journal of machine learning research* 13.Feb (2012), pp. 281–305.

52. Hofman Jake M., Sharma Amit, and Watts Duncan J. "Prediction and explanation in social systems". In: *Science* 355.6324 (2017), pp. 486–488. ISSN: 0036-8075. eprint: https://science.sciencemag.org/content/355/6324/486.full.pdf. URL: https://science.sciencemag.org/content/355/6324/486. PMID: 28154051

53. Raudenbush Stephen W and Bryk Anthony S. *Hierarchical linear models: Applications and data analysis methods*. Vol. 1. Sage, 2002.

54. Van Dusen Ben and Nissen Jayson. "Modernizing use of regression models in physics education research: A review of hierarchical linear modeling". In: *Physical Review Physics Education Research* 15.2 (2019), p. 020108. https://doi.org/10.1103/PhysRevPhysEducRes.15.020108

55. Rubin Donald B. "Inference and missing data". In: *Biometrika* 63.3 (1976), pp. 581–592. https://doi.org/10.1093/biomet/63.3.581

56. Cyril Goutte and Eric Gaussier. "A probabilistic interpretation of precision, recall and F-score, with implication for evaluation". In: *European Conference on Information Retrieval*. Springer. 2005, pp. 345–359.

57. Weiss Gary M, Kate McCarthy, and Zabar Bibi. "Cost-sensitive learning vs. sampling: Which is best for handling unbalanced classes with unequal error costs?" In: *Dmin* 7.35-41 (2007), p. 24.

58. Chawla Nitesh V et al. "SMOTE: synthetic minority over-sampling technique". In: *Journal of artificial intelligence research* 16 (2002), pp. 321–357. https://doi.org/10.1613/jair.953

59. Guyon Isabelle and Elisseeff Andre. "An introduction to feature extraction". In: *Feature extraction*. Springer, 2006, pp. 1–25. https://doi.org/10.1007/978-3-540-35488-8_1

60. Hosmer David W Jr, Lemeshow Stanley, and Sturdivant Rodney X. *Applied logistic regression*. Vol. 398. John Wiley & Sons, 2013. https://doi.org/10.1002/9781118548387

61. Didrik Nielsen. "Tree boosting with xgboost-why does xgboost win" every" machine learning competition?" MA thesis. NTNU, 2016.

62. Sterne Jonathan AC et al. "Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls". In: *Bmj* 338 (2009), b2393. https://doi.org/10.1136/bmj.b2393

63. Kezar Adrianna. "Higher education change and social networks: A review of research". In: *The Journal of Higher Education* 85.1 (2014), pp. 91–125. https://doi.org/10.1353/jhe.2014.0003

64. Bodong Chen and Oleksandra Poquet. "Socio-temporal dynamics in peer interaction events". In: *Proceedings of the Tenth International Conference on Learning Analytics & Knowledge*. 2020, pp. 203–208.

65. Vaupel James W, Manton Kenneth G, and Stallard Eric. "The impact of heterogeneity in individual frailty on the dynamics of mortality". In: *Demography* 16.3 (1979), pp. 439–454. https://doi.org/10.2307/2061224 PMID: 510638

66. Hougaard Philip. "Frailty models for survival data". In: *Lifetime data analysis* 1.3 (1995), pp. 255–273. https://doi.org/10.1007/BF00985760 PMID: 9385105

67. Byers González Julie M and DesJardins Stephen L. "Artificial neural networks: A new approach to predicting application behavior". In: *Research in Higher Education* 43.2 (2002), pp. 235–258. https://doi.org/10.1023/A:1014423925000