6-1-2018

Classification of One-Year Student Persistence: A Machine Learning Approach

Ben Siebrase
University of Denver

Classification of One-Year Student Persistence: A Machine Learning Approach

_____

A Dissertation

Presented to

the Faculty of the Morgridge College of Education

University of Denver

_____

In Partial Fulfillment

of the Requirements for the Degree

Doctor of Philosophy

_____

by

Ben Siebrase

June 2018

Advisor: Dr. Antonio Olmos

Author: Ben Siebrase
Title: Classification of One-Year Student Persistence: A Machine Learning Approach
Advisor: Dr. Antonio Olmos
Degree Date: June 2018

Abstract

Multilayer perceptron neural networks, Gaussian naïve Bayes, and logistic regression classifiers were compared when used to make early predictions regarding one-year college student persistence. Two iterations of each model were built, utilizing a grid search process within 10-fold cross-validation in order to tune model parameters for optimal performance on the classification metrics F-Beta and F-1. The results of logistic regression, the historically favored approach in the domain, were compared to the alternative approaches of multilayer perceptron and naïve Bayes based primarily on F-Beta and F-1 score performance on a hold-out dataset. A single logistic regression model was found to perform optimally on both F-1 and F-Beta. The logistic regression model outperformed all four of the individual alternative models on the evaluation criteria of concern. A majority voting ensemble and two additional ensembles with empirically derived weights were also applied to the hold-out set. The logistic regression model also outperformed all three ensemble models on the scoring metrics of concern. A visualization technique for comparing and summarizing case-level classifier performance was introduced. The features used in the modeling process comprised traditional and non-traditional elements.

Keywords: classification, machine learning, naïve Bayes, artificial neural network, multilayer perceptron, logistic regression, institutional research, persistence, retention, attrition

Acknowledgements

I would like to thank the University of Denver for allowing me the opportunity to study underneath the dedicated faculty of the Research Methods and Statistics Program. I would also like to extend a sincere thank you to Dr. Antonio Olmos for his continued support and guidance throughout my project. To Dr. Kathy Green and Dr. Shimelis Assefa, thank you both for your support and feedback throughout this process. I would also like to thank Dr. Janette Benson for her involvement in this project and for the mentorship that she provided me at the outset of my career. Also, thank you to my friend and colleague, Mike Furno, for encouraging me to pursue this project. Finally, a heartfelt thank you to my wife, Jamie, and my sons, Jon and Brian, for their constant encouragement, support, and inspiration.

Table of Contents

List of Tables

List of Figures

CHAPTER ONE: INTRODUCTION AND LITERATURE REVIEW

Problem Statement

Binary classification is a common task within many fields. The development, evaluation and comparison of different classifiers is a complex task requiring a high degree of standardization as well as the establishment of *a priori* evaluative metrics. Modern machine learning approaches to building, testing and selecting classifiers can be leveraged in order to optimize classification results for context-appropriate metrics. The current study will build and compare three different types of classifiers in order to identify the best performing approach within the context of college student persistence. Knowledge of the context within which the classification task is to be performed is critical to a successful modeling effort. The reviewed literature will begin with an overview of studies regarding college student persistence theory, moving through common approaches to classification within this field and then transitioning into relevant methodological research.

College student persistence has been an extensively researched subject over the past few decades, and efforts in this domain have ranged from theorizing cohesive theoretical frameworks for understanding persistence (Astin, 1984, Tinto, 1987) to using student-level data to make predictions about an individual's likelihood to persist. Persistence or retention, sometimes referred to inversely as attrition, have been operationalized in diverse ways with the timelines of concern fluctuating from one year

to four years to degree completion. Regardless of the approaches to persistence and the emergence of explanatory theories related to this phenomenon, the issue of persistence remains important and continues to have enormous financial, social, and reputational implications for educational institutions, students, potential students, families, and society in general (Carey, 2004; Ishitani, 2006; Kuh, Kinzie, Buckley, Bridges, & Hayek, 2007). Although persistence has been well-researched, there have been few successful efforts to leverage efficient and effective machine learning algorithms against unique and comprehensive student data stores in order to establish reliable and accurate classification models. Continued exploration of persistence using new, developing classification techniques is essential in order to establish and refine mechanisms for detecting and addressing issues with student persistence. Research of this type can lead to improved methods for identifying at-risk students in a timely fashion and should also enhance theory regarding the types of personal and environmental factors and contextual circumstances that have an impact on student persistence.

Research Purpose and Contribution

The current study sought to expand the body of literature regarding performance of classification techniques on student persistence by taking an applied approach to modeling the probability of persistence in six first-time first-year (FTFY) college cohorts at a private, non-profit, research University in the Rocky Mountain West. Specifically, this study makes a unique contribution to the current body of literature by utilizing techniques that have thus far been unconventional in the modeling of college persistence. Specifically, the study employed machine learning approaches, such as naïve Bayes and

artificial neural networks and compared those techniques against the most common analytic approach in this domain in an attempt to build an accurate classification model that could potentially be implemented in real time in order to provide what Tinto (1987) referred to as a "retention assessment system" (p. 191-203). The study also contributes to the extant research in this area by leveraging the increasingly large stores of somewhat obscure and unanalyzed student data in order to expand upon the set of traditionally considered persistence predictors.

A very limited number of studies, such as Delen (2012), have utilized neural networks and other less contextually traditional approaches to make college student persistence classifications and have compared the accuracy of those results against the results of logistic regression. The current study, however, departs from that research in several ways. First, naïve Bayes is employed in the current research and does not appear to have been used in a multi-analytic method comparative study on college student persistence. Second, the current study modeled the entire population of six cohorts of students instead of excluding entire groups of individuals. Third, as mentioned above, the data used in this research consisted of both traditional persistence predictors as well as less conventional and more emergent information such as transactional data generated from id card swipes, conduct violation records, and detailed information regarding housing, geographic location, and parking. Fourth, unsupervised methods were heavily leveraged in order to create unique and information rich composite predictors prior to the modeling phase. The creation of composite predictors not only offered insight regarding the formulae for generating useful condensed features, but also contributed to dimensionality reduction. Fifth, the evaluation of classification results in this study not

only examined sensitivity and overall accuracy. Instead, all relevant classification measures were compared, receiver operating characteristic curves were examined and a thorough descriptive analysis was conducted for correctly classified and misclassified results, in order to determine if certain models perform better within different subpopulations. Sixth, an ensemble model, which establishes a weighted average prediction for each case utilizing the outputs of all models was assembled and assessed. Finally, implicit differences in neural net architecture due to modeling decisions (i.e., activation function and number of hidden layers) resulted in a tailored, unique classification mechanism. The current study examined neural networks with hyperbolic tangent and sigmoid activation functions. Additionally, models were built with both one and two hidden layers.

The current study was applied and contains elements of replication and extension in regard to the small amount of work related to machine learning approaches to FTFY college student persistence in the literature. From an applied standpoint, the goal of the research was to build and identify the strongest model for classification within this very specific context. A byproduct of this process was an examination of the value in exploring and leveraging all available information about students in order to potentially discover new or poorly documented features with high predictive value. However, the research also served several second-order purposes. Specifically, the research was a general comparison study of machine learning versus logistic regression classification techniques with "real" data. The research also was one of only a very few efforts to compare these techniques within a college persistence context, and therefore has the potential to add further evidence regarding the efficacy of machine learning in the higher

education persistence domain. The research extends previous efforts in the field by incorporating ensemble modeling, specific neural net architecture, novel data, more rigorous feature creation, and a deeper approach to classification result evaluation. Specific research questions are presented below.

Research Questions

Research question 1a. Which of the examined models results in the most accurate classification of students into attrition and persistence groups?

Research question 1b. Does the ensemble model surpass the best performing individual model in terms of the examined classification metrics?

Research question 1c. Do any of the alternative algorithms perform better than the logistic regression model in regard to the metrics of concern?

Research question 2a. Which features or composite features result in the best performing classification model?

Research question 2b. Which of the rarely explored data elements are the most powerful predictors?

Research question 3a. To what extent do the models differ in terms of the cases that they accurately classify or misclassify?

Research question 3b. What is the profile of correctly classifiable versus incorrectly classifiable cases of attrition?

Research question four. Which neural network architecture results in the best classification performance?

Theoretical Framework

It is essential at the outset of this study to describe the context and theoretical framework within which the current inquiry was based. An institutional research perspective informed the majority of this study. Institutional research, specifically in the context of higher education, is a multi-faceted analytical discipline tasked with serving a particular institution or system with data-based decision making and analytical inquiry. In an analysis of prevalent and typical institutional research tasks in higher education, the Association for Institutional Research (2016) identified the following broad domains as essential to the field: accreditation, assessment, committee work, data integrity, data support, education, technology, planning, policy, program development, reporting, research, and student success. The research domain is arguably the most task-dense domain described in the report and lists functions such as analyses related to student outcomes, execution of ad-hoc research for institutional decision making, analyzing and reporting on admission, enrollment, and graduation trends, developing measurement instruments, conducting feasibility studies, statistically modeling various institutional outcomes, examining longitudinal processes, and executing research regarding student persistence (Association of Institutional Research).

As a profession, institutional research is characterized by elements of statistics, information management, programming, methodology, evaluation, and social science research. However, as a central part of institutional decision making, the field is clearly of an applied nature and can in many ways be seen as the intersection of the formal educational research tradition with business analysis characteristics typically associated with the private sector. The emphasis on application, or being able to leverage a

technique, method, or insight for organizational improvement or decision making, emerges from the institutional research perspective and was a substantial driver of the current research.

While the study and its development and justification considered theoretical aspects, the objective of the study was to achieve a practical and potentially implementable solution to a real problem. For example, the data utilized in the modeling were in part intended to reflect factors and attributes which have received theoretical or empirical support in the literature, and the strengths, weaknesses, assumptions, and contexts for successful application of the employed algorithms and methodological approaches are addressed. As specified above, the intent of the study was to show the utility of capitalizing on emerging and underutilized student data. Those data were selected with an exploratory intention, under the guidance of predominant college persistence theory. The contextual generalizability of powerful techniques that are unconventional to the discipline was assessed with the goal of achieving an accurate and actionable model. Therein, the area of contribution was primarily to applied educational research, specifically within the institutional research sphere.

Literature Review

In the following pages I present a review of relevant literature on the current topic, beginning with an overview of prevailing theories of student persistence, discussing studies that have empirically supported the role of specific factors in persistence and examining the practical utility of the primary methodological approaches utilized in the literature. I also discuss studies that have applied machine learning

techniques in an educational context and I then provide a more general discussion on machine learning techniques and their utility across a range of disciplines in order to further emphasize their potential to contribute to advances in the educational and social science realms. Finally, specific background on the relevant methods is provided.

Student Persistence Theory and Empirically Evidenced Predictors

Perhaps the most influential contribution to the study of student persistence has come from Tinto's (1986) unified theory of student departure from higher education, which emphasizes the importance of processes of academic and social integration in the individual student's successful retention. Much of Tinto's work and theory are based on aggregate observations across domestic institutions with varying levels of control, academic rigor, demographics, admittance standards, and academic outcomes. While the work is largely intended to be generalizable, Tinto notes that institution-specific research is the only means to gain knowledge regarding the nature of persistence at a particular institution. This observation is directly in support of the current research effort.

> In regard to the importance of social and academic integration Tinto (1986) notes: At the very outset, persistence in college requires individuals to adjust, both socially and intellectually, to the new and sometimes quite strange world of the college. Most persons, even the most able and socially mature, experience some difficulty in making that adjustment. (p. 47-48)

The above excerpt is broadly indicative of Tinto's theory. Tinto elaborates on this central notion by discussing the roles of what he calls intention, commitment, adjustment, difficulty, congruence, and isolation. These processes underscore the fact that Tinto views the dual integration processes as nuanced and multi-faceted.

Tinto (1986) also provides useful structural guidelines within which to consider persistence, noting that persistence can take the forms of "voluntary leaving" or "academic dismissal" (p. 83). This is an important distinction that Tinto deals with in more detail, but as one might intuitively surmise, there can be substantial fundamental differences between students who leave a college or university willingly and those who are forced to leave due to a failure to meet the minimum standards of academic performance. He also suggests that researchers, in an attempt to provide some level of standardization to the investigative processes, have traditionally focused on exploring persistence trends in cohorts of first-time first-year students. This is also a small but important observation which is reinforced in practice by mandatory institutional reporting such as the U.S. Department of Education's National Center for Education Statistics' Integrated Postsecondary Education Data System (IPEDS) reporting and voluntary rankings reporting such as US News and World Report's Best College Rankings. IPEDS defines persistence, or as they refer to it, retention rate, as:

> A measure of the rate at which students persist in their educational program at an institution, expressed as a percentage. For four-year institutions, this is the percentage of first-time bachelors (or equivalent) degree-seeking undergraduates from the previous fall who are again enrolled in the current fall. For all other institutions this is the percentage of first-time degree/certificate-seeking students from the previous fall who either re-enrolled or successfully completed their program by the current fall. (IPEDS, 2016).

The US News Best Colleges methodology for determining persistence is based on the definition provided by IPEDS and the importance of this metric in the third-party ranking of colleges and universities is evidenced by the fact that for their 2016 rankings, US News attributed 22.5% of an institution's overall score to retention as measured by the six-year graduation rate and the one-year persistence rate of first-time first-year

students (US News, 2016). As we see from the preceding references to current definitions and uses of persistence in commercial and federal reporting, it is essential to have strict definitions within which persistence can be explored. The temporal (one-year) and structural (FTFY) framework that Tinto called attention to in his early work are still one of the most common units of analysis for exploring persistence.

Tinto (1986) also offers information, to be discussed in later sections, on individual demographic attributes which have been shown to have associative or predictive relationships with persistence. However, these observations are largely consistent with much of the other literature in the field.

Astin (1984), a contemporary of Tinto, also constructed an overarching theory for college student persistence called Student Development Theory. Astin is critical of many of the predominant theories of student persistence, such as resource theory, which postulates that the greater the amount and quality of resources an institution possesses, the better their student outcomes are. Astin criticizes this particular theory for its lack of concern with the role of the individual student in the process of utilizing resources. Concern for the importance of the individual student is a hallmark of Astin's theory, which is highly student-centered. Astin describes student development as "the quantity and quality of the physical and psychological energy that students invest in the college experience," suggesting that higher levels of involvement in an individual lead to greater student learning and growth for that person (pp. 306-307). Astin goes on to note that all administrative policies and decisions should be directly targeted at increasing student involvement, which he further clarifies as being concerned with the "motivation and behavior of the student" (pp. 306-307).

One of the strengths of Astin's theory is its skepticism of unverified claims regarding mechanisms for improving student outcomes and its advocacy of empirically evidencing factors associated with student persistence. For example, Astin sites a number of attributes such as playing intercollegiate sports and living in a residence hall which have been shown in the literature to be associated with increased rates of persistence (p. 302). Astin suggests that in order to further refine Student Involvement Theory, there need to be increased efforts to identify and test varying types of involvement. The current research is carried out in the spirit of Astin's call for more rigorous quantitative exploration of characteristics, behaviors, and circumstances assumed to impact persistence.

Milem and Berger (1997) observe that both Tinto's and Astin's theories put emphasis on the importance of student behavior but that most efforts to test Tinto's longitudinal theory of social and academic integration employed respondent opinion instead of observable behavior. Milem and Berger acknowledge the appropriateness of using perception-based inquiry strategies to address Tinto's integration constructs but argue for a complementary melding of the theories of Astin and Tinto based, as the authors note, on their understanding of Walsh's (1973) suggestion that perceptions within a certain context lead to new behaviors within an individual which in turn lead to changes in the perceptions within the given context.

Milem and Berger's (1997) attempt to empirically and longitudinally integrate the two dominant theories in the field incorporate both fall and spring measures of observable behaviors as well as demographic characteristics, measures of perceptions of institutional and peer support, and measures of social and academic integration in order to

examine the impact on intent to reenroll after the end of the first year of college. Milem and Berger theorized that observable behaviors in the fall would impact perceptions of institutional and peer support which would then impact behaviors in the spring, which would then impact academic integration, commitment to the institution, and the intention to reenroll. The authors utilized a structural equation model estimating only direct effects in order to examine the relationships between behavior scales, perception scales, and the dependent variable.

The authors found, among many things, significant positive predictive relationships between peer involvement and the perceptions of institutional and peer support as well as between faculty interaction and the two aforementioned perception scales. The authors also found significant predictive relationships between academic non-engagement and perceptions of institutional support, where the less academically engaged the individuals the weaker their perceptions of institutional support. Non-engagement with the institution itself in the fall was also found to have a significant negative relationship with institutional commitment in the spring. Milem and Berger (1997) found perceptions of peer support in the fall to be predictive of social integration in the spring and also found fall perceptions of institutional support to be predictive of spring academic engagement. They also found that spring peer involvement had a positive predictive relationship with both social and academic integration as well as with institutional commitment. Finally, the authors found increased faculty interaction to predict higher academic integration. Notably, the site of Milem and Berger's research shares many characteristics with the institution serving as the site for the current research, potentially increasing the generalizability of some of the authors' findings.

In regard to the demographic characteristics of entering FTFY students, Milem and Berger (1997) found a direct effect in an SEM model between high school GPA and a measure of academic integration and also identified income level as a negative predictor of student levels of institutional commitment at the end of the first year of college. Astin (1984) mentions several factors as being positively associated with persistence. Specifically, he discusses living on campus, participation in honors programs, academic involvement, interaction with faculty, participation in student government, and playing intercollegiate athletics as having positive impacts on persistence.

The importance of interaction with faculty as predictor is found in a great deal of persistence research, and has been long held as playing an important role in students remaining at an institution. Pascarella and Terenzini (1979) provided some of the earliest empirical support for this notion by demonstrating significant increases in variance explained in student attrition models when including a measure of students' interaction with faculty outside of the classroom.

Milem and Berger's (1997) study, discussed above in terms of its results and theoretical contributions in regard to integrating the theories of Astin and Tinto, utilized factor analysis to derive scales of observable behaviors which were then utilized in the study's final path model. Milem and Berger's scales included involvement with faculty and peers, academic and institutional nonengagement, participation in social activities, participation in organized activities, and exercise/recreation. The retained items in Milem and Berger's scales included such behaviors as missing class, failing to submit work on time, participation in Greek life, consumption of alcohol, volunteering,

participation in student clubs and groups, participation in residence hall programs, and exercising at the institution's fitness center. Many of the resultant scales in Milem and Berger's study were found to have statistically significant predictive relationships with institutional commitment, with other scales' scores across time, with scales measuring perception of institutional/peer support, and with the endogenous variable itself. Notably, Milem and Berger's traditional social involvement scale, which consists only of three items and includes an item concerning the consumption of alcohol, is predictive of perceptions of peer support as well as social integration. However, this scale also has a negative predictive relationship with both academic and institutional nonengagement, suggesting that while behaviors such as social drinking might lead to greater social integration they may also lead to lower levels of academic integration.

The issue of alcohol consumption is addressed in greater detail by Martinez, Sher, and Wood (2008), in which the authors allude to the point made above, basing their research on the theory that participation in events such as, "Greek parties, intercollegiate sports events, and residence hall parties" is linked to social engagement and integration but that these types of events are also "strongly associated with heavy drinking" (p. 451). The authors examined the impact of heavy drinking while controlling for event attendance so that the impact of heavy drinking behaviors in excess of those exhibited at the relevant events could be examined. The authors identified a pattern in which sporting event, Greek party, and off-campus party attendance was positively associated with high levels of drinking and in which increased attendance at these events was predictive of increased rates of persistence. The authors also found that frequenting bars/clubs was predictive of high levels of drinking as well as higher rates of attrition. Attending parties

in dormitories was associated with less drinking but also predictive of lower persistence. The findings of Martinez are notable in that they draw attention to some of the complex processes and potential mediating relationships involved in the persistence process.

Financial factors have been extensively discussed in the literature on college student persistence as well. St. John, Cabrera, Nora, and Asker (2000) noted the potentially complicated impact of family resources and institutional aid, and highlighted that while contention exists in the literature regarding the exact roles of these attributes, they should be present in any persistence modeling effort. Braunstein, McGrath, and Pescatrice (2000) underscore the importance of examining family financial resources with their finding that students from families with greater resources were more likely to persist.

In their review on empirical persistence research, Ishler and Upcraft (2004) provide a broad list of the areas shown to have an impact on student persistence. The authors note the following important areas: prior academic achievement, socioeconomic resources, gender, age, race/ethnicity, parental support, student commitment, first-year GPA, field of study, enrollment status, quality of effort, faculty interaction, interpersonal interaction, extracurricular activities, work obligations, satisfaction, alcohol abuse, involvement in Greek organizations, perceptions of campus climate, financial aid, participation in athletics, classroom experiences, first-year seminars, orientations experiences, living quarters, learning communities, advising, service-learning, supplemental instruction, support services, and intervention such as faculty or academic advisor outreach. In a similar review of empirically evidenced persistence indicators, Therriault and Krivoshey (2014) echoed many of the areas cited by Ishler and Upcraft

and noted some specific precollege academic achievement areas such as minimum thresholds for high school mathematics credits, advanced placement exam scores, high school GPA minimums, SAT scores, and participation in dual enrollment programs during high school as significant predictors. The predictive usefulness of academic preparation is also supported by Stewart, Lim and Kim's (2015) finding that GPA score, ACT score, and receipt of remediation services were all significant predictors of persistence. Similar results regarding high school GPA, SAT score, and number of college credits earned before matriculation were found by Wu, Fletcher, and Olson (2008).

Gansemer-Topf, Zhang, Beatty, and Paja (2014), used a mixed-methods approach, combining a logistic regression analysis and interviews, in order to explore one-year persistence at what they note was a small, private, selective, liberal arts institution. The authors used demographic predictors as well as financial aid figures, admission characteristics, major, and GPA in their model, with only GPA emerging as a statistically significant predictor of one-year persistence. Specifically, the authors found that, on average, higher end-of-first-year GPAs were associated with persisting to the second year. Interestingly, the model correctly classified approximately 70% of the students with a correct classification rate of about 52% for the attrition group. The authors also note that three themes were generated from the qualitative component of their study, "struggling with college transition, realistic expectations of academic rigor, and social integration" (p. 276). These themes fit well within Tinto's (1987) model but when considered in the context of the results from the researchers' logistic regression model,

they suggest that there are some important and complex processes at play which are clearly not represented by the data used for modeling.

Notably, Gansemer-Topf et al. (2014) were primarily concerned with voluntary leaving and did not conduct interviews with those who were required to leave for academic reasons. This is an interesting distinction as it highlights two broad and fundamentally different classes of reasons for leaving college. The fact that these different reasons exist is intuitive, however, in a modeling context they may have important implications. For example, academic success, as measured by GPA, could reasonably be expected to have an important predictive relationship with one-year persistence. In fact, most institutions have a minimum GPA requirement that must be maintained, so if a student were to drop below that threshold there are points at which a predictive model could not be usefully applied to that student. Their GPA alone would guarantee dismissal. On the other hand, high performance should not necessarily be assumed to have a simple linear relationship with persistence, since student perception of insufficient academic rigor could also lead to departure. Regardless of the potentially complicated nature of GPA as a predictor of persistence, many studies highlight it as being significantly related to one-year persistence (Gansemer-Topf et al., 2014; Harvey & Luckman, 2014).

Within any discussion of modern persistence research it is important to note that the notion of one-year persistence of first-time first-year students, while likely the most common unit of analysis, is not the only way of considering persistence and retention. In recent years, more attention has been given to the complex and non-linear trajectories of students' college careers. McCormick (2003) calls attention to this in his work on the

ideas of swirling and double-dipping, concepts which essentially describe emerging patters of students alternating between institutions, bouncing back and forth between institutions or concurrently enrolling in multiple institutions. McCormick draws attention to these patterns in order to highlight the complexity of the student lifecycle and pathway towards degree, underscoring the importance of a student's ultimate outcomes regardless of what their outcome might be at a specific institution. For example, a student may transfer from one institution within their first year, return to that institution in their third year, and then depart from that institution again, take two years off, reenroll, and graduate from a third institution. Such a student would have achieved the goal of completing a degree, however, the outcome was not achieved at the initial institution and the student would not have been considered as having persisted for one year by their first school. These complexities must indeed be acknowledged as they have implications for how we define and study student success. They also draw attention to the fact, which will be discussed in more detail below, that the act of a student leaving an institution can be motivated by multiple and widely varying reasons. However, it must be acknowledged that the act of persisting within a single institution is an enormously important, nuanced concept that is not yet well enough understood.

A recent study by Campbell and Mislevy (2012), which was influenced by the notion of swirling, attempted to model multiple outcomes of college students at a single institution over the course of multiple years using the following predictors: engagement, academic ability, personal financial resources, belonging, educational aspirations, on-campus residency, race/ethnicity, and gender. The data utilized in this study were self-reported survey data derived from the responses of first-time first-year students to a

preexisting instrument. The researchers explored the outcomes of continuous enrollment, stop-out, drop-out, and transfer using multinomial logistic regression. Notably, Campbell and Mislevy's data were based on perception and not observed behaviors or characteristics.

Campbell and Mislevy (2012) also highlight another important characteristic of persistence studies in that it uses logistic regression or multinomial logistic regression to classify students into various categories of persistence outcomes. In a classification study, oftentimes the objective is to identify at-risk students so that an intervention can potentially be developed or administered. The success of such studies can be misleading depending on the context of the study. For example, at a school with a relatively high persistence rate, attrition is a rare outcome. Since most of the students at such a school will be expected to persist, classification tables will likely demonstrate great success in classifying the most common outcome. However, the rarer outcome, the outcome of concern, in this case attrition, will be characterized by a low success rate. The average classification, however, can be relatively high, suggesting an effective model. In the case of Campbell and Mislevy (2012) the researchers successfully classified only 2% of stop-outs and 9% of transfer-outs for females but achieved a 72% overall classification accuracy rate for this group. The researchers overall successful classification rate for males was even higher at 83%, however, the model classified all participants into the continuously enrolled group, failing to accurately or inaccurately detect any potential students at risk of leaving. Models that are not sensitive enough to detect the outcome of concern have limited practical utility. Notably, the researchers highlighted the limitation of not having individual-level classifications expressed in an actual probability.

Primary Methodological Approaches to Persistence and Machine Learning Efforts in an Educational Context

As the preceding information suggests, attempts to predictively model persistence, especially when the outcome is dichotomized, have historically favored the use of logistic regression (Gansemer-Topf, et al., 2014; Glynn, Sauer, & Miller, 2006; Lopez-Wagner, Campbell, & Mislevy, 2012; Miller (1999); Robb, Moody, & Abdel-Ghany, 2012). This is not to say that logistic regression is a poor choice of model. Indeed, given the disciplinary context within which these studies are typically conducted, it is to be expected that logistic regression is frequently utilized. Moreover, logistic regression should not necessarily be assumed to be an outdated or underperforming model, as it is still a very flexible, powerful, and oft used model which is generally grouped under the umbrella of machine learning classification algorithms. Logistic regression is often utilized in a formulaic sense for explanatory purposes in which variable significance is assessed. However, it is easily applied to contexts where prediction, not explanation, is the primary concern. Lower performance of some previous one-year persistence classification efforts using logistic regression should not necessarily attribute classification performance to the inadequacy of the model but instead to the complexity of the modeling scenario. However, as new and sophisticated algorithms have emerged and been used successfully in other disciplines, it is important to evaluate the efficacy of these algorithms at modeling phenomenon across other disciplines, such as college student persistence.

The potential for alternative algorithms to outperform the standard logistic regression model are abundantly represented in the literature. In a comparison of 72

medical studies using both logistic regression and artificial neural networks for comparative purposes, Dreiseitl and Ohno-Machado (2003) found that neural networks performed better than logistic regression models on the same data in 51% of reviewed studies, whereas there were no performance differences on 42% of the cases and logistic regression outperformed neural networks on only 7% of studies. The authors also note the comparative ease of building and interpreting logistic regression models. It should be noted that other more traditional modeling techniques are also seen in the literature, such as the use of structural equation modeling by Cabrera, Nora, and Castaneda (1993) and Milem and Berger (1997). However, these approaches are seen at a much lower rate and have been less prominent in recent years.

In an applied study utilizing logistic regression for classification, Miller and Tyress (2009) found several demographic factors as well as high school GPA and responses to several questions about expectations regarding the college experience to be predictive of persistence. Notably, the model correctly classified only 16.79% of students who attrited, while misclassifying 10.99% of students who were retained. Miller and Tyress highlight the use of data gathered before matriculation as a benefit of the study. Indeed, utilizing data that can be generated before the college experience has value, although the extent to which the predictive value of these data generalize across specific institutional contexts is uncertain, and it is likely that more precise predictions can be made by incorporating data generated from the actual college experience.

In an applied institutional research project in the California State University system, Lopez-Carollo, and Shindlecdecker achieved 80% correct classification of FTFY attrited students and an overall correct classification rate of 77% using a logistic

regression model. The model examined demographic variables, high school GPA, and various enrollment variables, such as participation in a specific curriculum and proportion of core courses completed during the first year. The overall model as well as several of the predictors emerged as statistically significant. While the overall classification rate as well as classification of the group of concern were fairly high, it should be noted that the model yielded a low level of precision with only 34% of the predicted attrition cases being accurate. This observation underscores the importance of holistically assessing classification results when evaluating a supervised model of this type.

Additionally, Lopez-Wagner et al. utilize data from the end of the first academic year in order to make predictions about an outcome which, at that point, is relatively near to manifesting itself. In an applied context, where it is desirable to use the results of the model to target potential interventions to those who are deemed most likely to attrite, models which can make accurate predictions earlier in the year are more valuable as they allow for a much larger window within which interventions can be administered. Notably, Lopez-Wagner et al. extended their findings in a practical manner by using the resulting model to generate individual-level probability scores and then transforming those scores into deciles. The authors then gained support for the generalizability of their model to future cohorts by testing its efficacy with another cohort.

Very few studies in the domain of college student persistence have utilized methodological approaches similar to those proposed for the current study. However, such studies are indeed beginning to emerge in the applied literature on this topic. For example, Borkar and Rajeswari (2014) used association rules and neural networks to model a measure of university performance in India.

The most recent and most methodologically similar study on student retention by Delen (2012) examined one-year persistence of several cohorts of first-year students. Delen made a substantial departure from the predominant methodological approaches to persistence research by utilizing neural networks and decision trees in addition to the more common logistic regression model. Delen also noted the practical implications of attempting to make predictions towards the end of the first semester in order to allow for an intervention period. Similar to the current research, Delen utilized predictors that were grounded in the literature and theory of college student persistence but also created composite features in order to potentially enhance the models. This process of feature creation was intentionally exploratory and was directly in line with Delen's data mining framework in which the potential for new knowledge discovery through the application of algorithms is tantamount to using algorithms to evidence *a priori* hypotheses.

Delen (2012) directly acknowledged the notion of advancing and augmenting theory as well as working towards practical applicability through this comparative study in which the efficacy of newer algorithms were compared against more traditional approaches. Delen's results also align with much of the literature regarding classification algorithm performance comparisons, as neural networks and decision trees both outperformed logistic regression models.

Delen (2012) achieved an overall classification rate of 81.19% with the neural network, classifying 93.83% of the attrition group members correctly but only 68.55% of the non-attrition group correctly. The logistic model resulted in an overall rate of 74.33%, with 85.34% of attriters classified correctly and only 63.31% of persisters classified accurately. The decision trees also performed moderately well overall with a

total classification rate of 78.25%, although, as with the neural network, the decision tree resulted in a very high rate of accurate classification for attriters, 92.53%, and a somewhat poor rate of correct classification for the persistence group, 63.96%. As noted above, it is rare to see such high rates of correct classification for an attrition group, especially in four-year colleges with moderate rates of attrition. The site for Delen's research had an eight-year average one-year attrition rate of 21.03%, which is moderately high, but still a relatively small group overall. This high rate of accurate classification for the group of concern must be considered alongside the accuracy rates for the persistence group, which suggest a high rate of false positives. Delen argues, however, that in cases where an intervention may be given, it may be acceptable to have a high false positive rate as long as there is sufficient discrimination between classes and there is a high rate of classification for the group of concern.

The current research attempts to both extend and replicate Delen's (2012) research. Delen made predictions about one-year persistence after the first semester, but the current research made predictions at the end of the first quarter, a slightly earlier time point. The current research also examined some additional classification techniques and gave strong consideration to preprocessing, parameter optimization, and the efficacy of ensemble techniques. The current research also utilized different and novel student predictors within a unique institutional setting featuring a considerably higher rate of persistence.

Machine Learning Applications and Successes in Non-Educational Contexts

The literature from computing and other scientific disciplines is replete with examples of the efficacy of supervised machine learning classification techniques. Supervised learning is a prevalent type of modeling where a model is built using known cases with relevant predictors and known outcomes (Marland, 2014). In these efforts, the goal is to perform estimation or classification on a known endogenous variable, as opposed to unsupervised techniques such as clustering, where the goal is to use input features to form empirical groupings that can be interpreted *post-hoc*.

Due to the infrequency of their use, there are relatively few examples of machine learning techniques outperforming traditional methodologies in an educational context. However, innumerable examples have emerged over the past twenty to thirty years in other disciplines. For example, Hepner (1990) employed a traditional classification technique for land-cover satellite photos with a large dataset and compared those results to those obtained from a neural network employing a quarter of the data and obtained similarly accurate results. The author also found that the traditional technique greatly underperformed compared to the neural network when using the smaller dataset. In an empirical examination of predictive performance of several popular algorithms, Caruana and Niculescu-Mizil (2006) found that artificial neural networks greatly outperformed both naïve Bayes and logistic regression, with logistic regression achieving a slightly higher accuracy rate than naïve Bayes in the given domain.

Broad Analytic Techniques

As the above literature suggests, predictive modeling efforts related to persistence are dominated by the use of classification techniques, the broad family of algorithms used to predict categorical outcomes. The current study, as well as much of the cited literature on persistence, utilize a simple binary outcome variable with two possible values, persist or attrite. According to Pereira, Mitchell, and Botvinick (2009), a classifier essentially finds the relationship between a set of features or predictor variables and the value of concern for a categorical outcome. Kohavi (1995) described a classifier as, "a function that maps an unlabeled instance to a label using internal data structures. An inducer, or an induction algorithm, builds a classifier from a given dataset." (p. 1). In this sense, classification is the result of applying a classifier, which is a function built from an inducer or a classification algorithm. Following is a discussion of the three classification techniques that were used to model the data in the current research.

Before discussing specific techniques and the study's methodology, additional consideration must be given to the methodological goals of the study in order to disambiguate the study's intentions. The current research compared classification techniques while building the most accurate classification model for the given data. Some of the techniques examined, especially logistic regression, are strongly associated with the explanatory research tradition. The current study, however, does not seek to explain the underlying mechanisms of attrition but instead to predict the attrition outcome. Sainani (2014a) offers insights on the differences of explanatory and predictive modeling, noting that explanatory research is intended to identify causal relationships, identify confounding variables, and test theoretical assumptions using smaller and

purposefully selected variables while the intent of predictive research is to accurately classify or estimate outcomes and build generalizable models for diagnostic purposes. Sainani also notes that model coefficients and statistical significance are critical to explanatory research while overall accuracy are the primary concerns of predictive research. The author also discusses the importance of training and testing predictive models on different datasets or data subsets. Sainani also adds that it is common to use larger, more exploratory datasets as well as dimensionality reduction strategies such as principal components analysis in predictive studies. These observations are important to note at the outset of the research as they help highlight the distinction between the two types of research and justify some of the modeling decisions and evaluative criteria used in the current predictive modeling study.

Logistic regression.

Binary logistic regression is similar to ordinary least squares regression, except that instead of using independent variables or features to estimate a continuous endogenous variable, the goal is to estimate the positively and negatively unbounded log odds, also called the logit, of being classified into one of two categories. When approaching a classification problem where the outcome variable, Y, represents the presence (Y==1) or absence (Y=0) of some condition, the logistic regression model predicting the log odds of Y being equal to one and utilizing predictors X_1 through X_j is as follows (Menard, 2013). In Equation 1, β_0 represents the constant and β_1 through β_j represent the model weights associated with each predictor X_1 to X_j.

$$Logit(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \ldots + \beta_j X_j \quad (1)$$

The odds that Y = 1 in this case can be derived by the following exponentiation of the logit (Menard, 2013).

$$odds(Y=1) = e^{\beta 0 + \beta 1\, X1 + \beta 2\, X2 + \beta 3\, X3 + \ldots + \beta j\, Xj} \quad (2)$$

Additionally, the odds of Y = 1 can be converted into the probability that Y = 1 by dividing the odds of Y = 1 by one plus the odds of Y = 1 (Menard, 2013). Abbott notes that the probability form of the equation is known as the logistic curve and results in values bound by 0 and 1 (p. 232). Essentially, the logistic curve transforms the results of the linear equation, for which the dependent variable is unbound, to a probabilistic function for determination of class membership. While classification tables are useful for assessing the results of any classification model, Sainani (2014b) notes that the receiver operating characteristic (ROC) curve, which plots model sensitivity and specificity for various class membership probability estimates, is also common to logistic regression and can serve as a useful mechanism for assessing model effectiveness. Sainani also emphasizes using caution while interpreting odds ratios, as variable magnitude and scaling considerations must be taken into consideration during interpretation.

It must be noted that the current study sought to compare more modern machine learning techniques with logistic regression, which has been shown to be the historically favored method of classification in the current context. However, the term machine learning is used broadly, and while the alternative techniques of neural networks and naïve Bayes are more commonly associated with machine learning efforts, logistic

regression too is often categorized under this umbrella. Good justification for this is provided through Menard's (2003) discussion of logistic regression estimation, where the author notes that maximum likelihood is used to estimate the log-liklihood function through an optimization process as opposed to the ordinary least squares technique used in linear regression. This repetitive process of searching for solutions in the data and adjusting coefficients in order to optimize a metric or cost function is a hallmark of machine learning. The overlap between logistic regression and neural networks is highlighted by Dreiseitl and Ohno-Machado (2002), as the authors note that both use optimization techniques in model building and that a neural network employing a sigmoid activation function without a hidden layer is equivalent to a logistic regression model. The authors also note the straightforward nature of interpreting logistic regression coefficients compared to neural network models and suggest there is a risk of overfitting with neural networks but also the potential benefit of deeper learning when compared to logistic regression.

In regard to model evaluation, Abbott (2014) notes that when building a logistic regression model, predictor significance can be used as a criteria to drop predictors, but that if classification accuracy is the goal, then it is unnecessary to eliminate non-significant features, especially if they enhance classification accuracy (p.235).

Naïve Bayes.

Naïve Bayes is a relatively simple classification approach based on Bayes' Theorem, which can be expressed as follows.

$$P(A|B) = \frac{P(B|A) * P(A)}{P(B)} \quad (3)$$

Downey (2013), provides what he calls a diachronic interpretation of this theorem where A is the outcome of concern or the hypothesis and B represents the observed data. In this interpretation, the term P(A|B) is known as the posterior distribution and can be interpreted as the probability of the outcome of concern given a set of features or attributes. P(B|A) is known as the likelihood and is the probability of observing the data given the outcome of concern. P(A) is known as the prior and is the probability of the outcome without regard to any data. P(B) is known as the normalizing constant and is simply the probability of a certain set of features regardless of the outcome of concern.

Bayesian networks are sophisticated extensions of Bayes' Theorem that model the dependencies between the conditional probabilities of data features. Naïve Bayes is essentially a Bayesian network that assumes independence between the input variables (Fan & Poh, 2009). Shmueli, Patel, and Peter (2016) note that in naïve Bayes the probability of a set of features conditional on a specific outcome is calculated by taking the conjoint conditional probability of those features, or in other words, by multiplying all of the individual conditional feature probabilities by one another. The authors note that this technique is often sufficient for classification but that it results in model probabilities that do not actually reflect the probabilities observed in practice for the modeling context. Rish (2001) observes that the assumption of independence between predictors is often unrealistic but that naïve Bayes classifiers have nonetheless been proven to be very effective at classification. Rish also finds, as one might expect given

the assumption of input independence, that naïve Bayes performs best when features are actually truly independent of one another. This observation is reflected in specific preprocessing steps that are suggested for naïve Bayes, such as the application of principal component analysis before modeling in order to build composite features and reduce dependencies between model inputs (Fan & Poh, 2009). Finally, in a comparison of naïve Bayes and logistic regression error rates using simulated data, Ng and Jordan (2002) find that naïve Bayes will typically feature a lower initial error rate but that logistic regression has a tendency to converge upon and even drop below those observed error rates given a large enough dataset.

Artificial neural network.

Artificial neural networks (ANNs) are a type of machine learning algorithm based on human brain functioning which are capable of modeling complex linear patterns and can be used for estimation as well as for classification (Shanmuganathan, 2016). The ANN has many applications and derivatives. One of the most popular forms of an ANN is the multi-layer perceptron (MLP) which, as Du and Swamy (2014) note, is a feedforward network with no connections between neurons within the same layer, no feedback between layers, and with every neuron connected to each neuron in subsequent layers.

According to Abbott (2014), the MLP ANN comprises layers of connected perceptrons, which are single-neuron ANNs, and for which all input variables receive a weight, forming a linear function. Abbott notes that an activation function, sometimes referred to as a squashing function, is applied to the sum of the weighted inputs for each

neuron. A common squashing function is the sigmoid, a continuous and non-linear function which takes and transforms the combined linear output of the weighted inputs. Abbott notes that the continuous nature of the sigmoid allows for the estimation of derivatives and the non-linear nature of the function is what allows ANNs to learn complicated, non-linear relationships. Marsland (2014) echoes this argument, noting that linear functions can only model problems with linearly separable classes, that is, groupings that can be separated by a hyperplane. The hyperbolic tangent activation function is also a popular alternative to sigmoid, though not as prevalent in the literature. In an experiment where multiple activation functions were compared with an otherwise identical architecture, Karlik and Olgac (2010) found slightly higher classification accuracy with a hyperbolic tangent function than with sigmoidal activation functions.

The basic MLP shown in Figure 1 has inputs or predictors X1 to XJ in the first layer. Each input is used as a predictor in each of the 1 to J linear functions in the hidden layer. The squashing function is applied within the hidden layer and the transformed results for each of the hidden perceptrons are then used to predict the final output values. This is the essence of the basic MLP, aside from the critical process of backpropagation, which is described below and not represented in the figure. In a supervised learning application, such as the one presented in Figure 1, the model is trained to the point that it can represent within its internal architecture the association between all X model inputs and their associated Y outputs (Shanmuganathan, 2016).

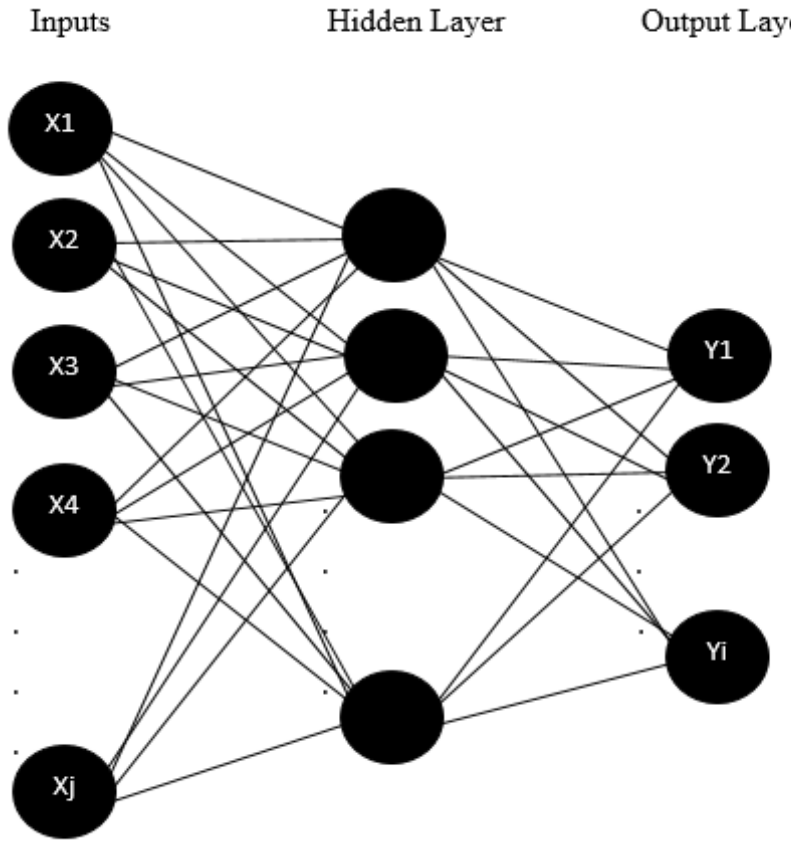


Figure 9. Multilayer Perceptron Artificial Neural Network with One Hidden Layer

Among the observed advantages of the MLP are the ability to model non-linear relationships. However, noted disadvantages include the fact that MLP features a non-convex loss function (exposing it to the risk of identifying local minimums during model optimization), the amount of parameters that need to be set and tuned during model building, and sensitivity to model input magnitudes based on normalization techniques or a lack thereof (Scikit-Learn Documentation). LeCun, however, notes that machine learning has been stifled by hesitancy to utilize non-convex optimization approaches, noting that convex approaches such as logistic regression are prevalent but that non-

convex approaches can be beneficial as they often allow for deeper learning and more detailed findings.

Backpropagation is the critical process by which neural networks estimate error during training and then adjust the weights from the inputs to the hidden layer neurons in order to minimize cost. Riedmiller (1994) notes that MLPs are essentially optimization problems with a goal of minimization, where the unit to be minimized across training is error, or the difference between known and predicted training classes. The mechanism for this optimization process is backpropagation, and the most common method for implementing these adjustments is gradient descent (Riedmiller, 1994). The implementation of backpropagation via gradient descent is described by Riedmiller as taking output values and then successively calculating the derivatives of the neurons in the layers before those outputs. After calculating derivatives, a small-scaled adjustment is made to the weights from inputs to the hidden layer by multiplying the negative derivative by the analyst-defined constant learning rate in order to move down the gradient and minimize error or cost. Typically, these backpropagations of error are made after a single modeling of the data using the entire training set. One pass through the data is known as an epoch. This technique is computationally simpler as well as quicker than backpropogating the errors after each case is modeled, and it has been used frequently employed since Rumelhard, Hinton, and Williams (1986) proposed gradient descent-based backpropagation in MLPs.

As mentioned briefly above, MLPs are characterized by a variety of training considerations that must be made. The backpropagation clearly relies on passing the training data through the model multiple times, but the ideal number of passes, or epochs,

is a decision that can be made and experimented with during model building. Gradient descent learning rate is also a consideration, as larger learning rates will result in quicker model computation, but excessively large or small rates could cause a model to pass over a global minimum or to identify a local minimum. Network architecture, such as the number of hidden layers and the number of neurons in each layer, is also important. MLPs are initiated by using random seeds to serve as initial synapse weights, and these weights can also potentially impact model results.

Pereira et al. (2009) note that in certain contexts, linear classifiers such as logistic regression are preferable to more complex non-linear models such as ANNs, as the simpler linear models, under the right circumstances, can generate comparable accuracy and are less complicated to interpret.

Feature selection.

A discussion of modern machine learning techniques for classification also highlights the need for discussing variable selection or dimensionality reduction strategies. Pereira et al. (2009) note the importance of being intentional about the number of variables considered in a modeling effort and advocate either or both selecting individual features purposefully or implementing dimensionality reduction strategies, such as principal component analysis in order to create composite features. The authors also note the importance of normalizing data during a preprocessing phase so that variables measured by different scales or with naturally larger magnitudes do not have an undue impact during training.

Guyon and Elisseeff (2003) also advocate dimensionality reduction and feature creation strategies as a preprocessing step in the model building process. The authors recommend k-means clustering, a process for partitioning cases into k number of clusters based on the case attributes. In this process groups with similar feature vectors are partitioned and those group memberships can serve as predictors. This particular clustering technique has been employed recently in order to enhance student success modeling in an educational context (Dutt, Aghabozrgi, Ismail, & Mahroeian (2015). Guyon and Elisseeff (2003) also discuss the concepts of saliency, entropy and density suggesting that these qualities can be used to generally distinguish the quality of potential features. Salience refers to high-variance features. Entropy refers to uniformity of distribution. Density refers to multicollinearity of variables. The authors also suggest utilizing a variable ranking process where variables are evaluated based on their individual predictive power or their correlation with the outcome of concern.

Principal component analysis.

Principal component analysis (PCA) was considered as a means to create information-rich features while simultaneously reducing the dimensionality of the data. Kambhatla and Leen (1997) note that dimensionality reduction is a critical step in data preparation for classification tasks and explain that the underlying goal is to obtain parsimony in the data through the establishment of condensed features that accurately capture the nature of the information. Cao, Chua, Chong, Lee, and Gu (2003) explain that PCA utilizes the covariance matrix of any matrix X, or the product of X^T and X in order to obtain a series of eigenvectors, also known as principal components, ordered by

eigenvalue, where larger eigenvalues explain greater variance in the data. The authors note that the transformations to the original matrix vectors of X are linear and that components of each eigenvector are computed orthogonally. A subset of eigenvectors can then be selected in order to explain the maximum variation in the data in lower dimensional terms.

Balancing.

The problem of unbalanced data is a well-known problem in the domain of machine learning that arises when an analyst is attempting to build a model in order to predict a rare or infrequently occurring class (Weiss, 2004). Oftentimes, the data available for modeling will reflect the rarity of the class in that the majority class will occur much more frequently, making it difficult to train a model that can effectively distinguish the rarer class, the class of concern. Weiss notes that techniques such as clustering can help alleviate the impact of rare cases, but also emphasizes a concept addressed previously, that is, that overall accuracy is not the most appropriate metric for evaluating classification models involving rare outcomes. Weiss suggests several potential strategies for dealing with class imbalance in a dataset, noting that under- and over-sampling are often used in order to, respectively, reduce the amount of majority cases in the dataset or to increase the number of minority cases. Weiss notes that both techniques have potential drawbacks, suggesting that over-sampling often involves making exact replications of minority class cases which can often lead to over-fit models, while under-sampling requires elimination of data that could potentially be useful.

The problem of imbalanced cases is present in the current research, and its magnitude will be discussed in the following section. In order to preserve as much of the data as possible for the current project, an over-sampling approach known as synthetic minority over-sampling technique (SMOTE) will be used to generate synthetic minority cases based on the nearest neighbors of the sampled minority cases. The application of this technique has been shown to result in more generalizable models by reducing the overfitting associated with simple replication of minority class cases (Chawla, Bowyer, Hall, & Kegelmeyer, 2002).

Glossary.

The remaining chapters will use several technical terms to describe the current research. Following is a description of key technical terms used in the remaining text.

Accuracy: In reference to classification results, accuracy refers to the proportion of all cases that are classified correctly.

Ensemble: An ensemble refers to a type of model that combines the predictions from multiple models to deliver a final prediction. Ensemble models have the potential advantage of incorporating the differential predictive value of various models in order to generate enhanced complimentary models.

F1: F1 is the harmonic mean of precision and recall. This metric takes into account both the proportion of accurately identified cases of concern as well as the proportion of predicted cases of concern that were correctly classified.

F-Beta: F-Beta is a weighted harmonic mean of precision and recall where the parameter, Beta, represents the weight assigned to precision in the computation of the

weighted mean. Beta values of less than one assign more importance to precision, while Beta values greater than one assign more importance to recall. During the remainder of the study, F-Beta will be used to refer to an F-Beta score where Beta equals two.

F1-Optimized: F1-optimized will refer to a specific model whose parameters resulted in the most desirable average F1 score during cross-validation.

F-Beta-Optimized: F-Beta-optimized will refer to a specific model whose parameters resulted in the most desirable average F-Beta score during cross-validation.

Grid Search: Grid search is a process implemented within cross-validation wherein which several potential values are given for a selection of model parameters. All permutations of parameter values are then assessed within cross-validation in order to identify the set of parameters that result in the best average scoring metric for which the model is being optimized.

Precision: Precision is an important scoring metric in the current study. This term describes the proportion of accurately predicted cases of attrition out of all predicted cases of attrition. Essentially, precision describes how precise model predictions were on the cases of concern. Precision values range from 0 to 1, with 1 representing perfect precision.

Recall: Recall is the proportion of correctly predicted cases of attrition out of all actual attrition cases. Recall is also referred to as sensitivity and is an important scoring metric within the current study, both on its own and in regard to its roll in F1 and F1-Beta. Recall essentially describes the extent to which a model is able to correctly identify the cases with the outcome of concern.

ROC: As described above, ROC stands for receiver operating characteristics. ROC curves are plots of recall against the false positive rate for a model along various decision boundary thresholds. All binary classification models evaluated in the current study offer probability scores for each outcome. The binary classification is then created using these scores and a constant decision threshold. However, the decision threshold can be altered. ROC curves show the impact of altering decision thresholds in terms of recall and the resulting false positive rate.

ROC AUC: ROC AUC represents the area under the curve, and refers to the size of the space beneath a specific ROC curve. AUC ranges from 0 to 1, with 1 representing a perfect classifier and 0.50 representing a completely random classifier.

CHAPTER TWO: METHOD

Design

The current research design utilizes historical institutional data in order to examine the impact of various student experiences and characteristics on one-year persistence with the ultimate goal of building an effective classification model for use at an early detection point during the first year of college. The study is influenced by the predominant theories of persistence and in that regard has a confirmatory component. The research setting and the use of novel variables and analytic techniques that have been seldom applied in this context also lend an exploratory element to the current research.

After a descriptive exploration of the data, the general progression of the current project consisted of a process of evaluating missingness and assessing correlation matrices and item distributions. Z-score normalization, dimensionality reduction, feature creation, and variable selection through k-means clustering and principal component analysis were done within the k-folds training and testing process in order to prevent leakage between the training and testing sets during each iteration. Multi-layer perceptron ANN, naïve Bayes, and logistic regression models were trained and tested during the k-folds cross-validation process and their classification results evaluated. Four permutations of ANNs were built, varying the number of hidden layers between one and two and varying the activation function between sigmoid and hyperbolic tangent. Between-model attributes and effectiveness were then compared. Finally, the utility of

an ensemble model consisting of all or some of the initial models was evaluated. All models were trained to predict the binary outcome of attrit or persist at the point of the third week of the second year of enrollment. Models were built from data collected prior to matriculation as well as during and up to the end of the first term of enrollment.

As mentioned earlier, the goal of developing an accurate classification model for first-year persistence is to be able to deliver appropriate interventions in order to increase the likelihood of retaining students. In that vein, early detection of students who are likely to leave the university is essential, and the first and earliest time point for which a model was developed (end of first quarter) received primacy.

The model evaluation process was also comparative and diagnostic in nature and evaluated the most effective models in terms of several classification table indices. The goal in this stage was to assess the nature of misclassifications, especially false negatives, in order to theorize data needs for future research efforts. In-depth descriptive analysis of accurately and inaccurately classified cases was also evaluated in order to determine the extent to which certain models are more or less effective within different subpopulations and to assess the profiles of correctly and incorrectly classified groups.

Participants

The current study initially considered the entire population of FTFY college students over a recent six-year period at a private, research university in the Western United States with an approximate enrollment of 10,000 students. Specifically, the study utilized cohort data from the falls of 2011 through 2016. The detailed IPEDS definition of an FTFY student can be found in Appendix A. The process for establishing cohorts is

automated and standardized by the institution and occurs at the end of the third week of the fall term. FTFY students still enrolled at this point in the term are assigned an attribute to denote their membership in the fall cohort of FTFY students. This attribute is from then on a part of each student's record and can be used to establish persistence and graduation rates at the cohort level.

Procedure

The institution utilizes a data freeze process in which data are frozen and archived during the third week (WK3) and end of the 10-week term (EOT). In order to convert the data to an optimal format for modeling purposes, a dataset was created where each individual is represented by a single row of data. To accomplish this a quarter/census indicator was created for each record by simply concatenating the quarter and census values for each record. A file consisting of only the variables ID, term, quarter, census, quarter_census, and WK3_cohort was then pivoted using the ID as index and quarter_census as the columns. Binary indicators were then imputed based on whether or not valid records indicating enrollment were present for each student at the various quarter_census points. The resulting dataset served as the base data used to indicate persistence. Valid enrollment records for the fifth quarter at WK3 served as the outcome variable, as they demonstrate whether or not students had met the threshold needed to establish one-year persistence.

The data for the current analysis was sourced from multiple institutional repositories. The core data, providing the outcome variables, were collected from a database containing frozen, census-bound historical student academic records.

Additional data were extracted from less-frequently utilized but centrally maintained tables in the institutional data warehouse. Information not historically archived or maintained by the institution, but instead hosted off-site, were also accessed. Data of this type include student transactional data generated from card swipes as well as conduct records and faculty feedback.

The overall dataset comprised several broad categories of information established based on extant literature and practical intuition. Sources of information with little empirical link to persistence but that have the potential to represent student behavior or to serve as proxies for engagement were included. These general areas included admissions qualifications, previously completed credit, financial aid, conduct records, institutional participation, academic performance, housing, parking, geographical location, curriculum, and demographics. At an even broader level, the intention of these areas was to capture a proxy for academic and social integration as well as ability and ambition.

The final dataset used in the model building process contained 115 variables. One or more measures were obtained for some variables depending on the context and nature of the variable. Measures for some variables were obtained quarterly at either the WK3 or EOT freeze point. After dummy coding and feature transformations the initial dataset contained 279 data points for each subject. See Appendix A for a list of all variables, their brief descriptions, and the number and names of the associated measures.

Preprocessing: Dimensionality reduction, composite variable creation and variable selection

The initial dataset was large and consisted of all variables with an intuitive, theoretical, or potential link to FTFY persistence. In order to maximize the predictive

usefulness of each data element, some variables were combined into composites. For example, low variance binary features indicating certain types of conduct violations were grouped together to form a binary indicator for a broader type of violation or, alternatively, violation indicators could have been summed to form ongoing violation counts within a specific area. The potential to reduce dimensionality with principal component analysis (PCA) in order to compute composite factor scores consisting of several similarly loading features was evaluated. The impact of PCA on cross-validation testing scores was evaluated against the select K-best univariate feature selection method. Many of the individual features had limited or no use in the modeling processes, and were therefore dropped from consideration. Specifically, low-variance or highly multicollinear features with no role in a composite variable were discarded. There is a potential for separate models to favor different features in the prediction of one-year persistence, so an initial set of potential features was established and then recursive feature elimination was performed for each model during the model building phases in order to select the best subset of predictors for the specific model and time point. All data were normalized or rescaled.

Analysis

The intent of the current study was to both build the most accurate classification model (or ensemble of classification models) using a wide range of student data and to also evaluate the classification accuracy of more emergent algorithms versus the more traditional classification technique in this field, logistic regression. The alternative modeling techniques tested against logistic regression were MLP ANN (four architectural permutations) and Gaussian naïve Bayes. The data available up to the time threshold was

modeled a number of times using each technique through a grid search process. Grid search allows for the specification of multiple potential values for any model parameter. Each permutation of model parameters is then assessed during the cross-validation process, allowing for the empirical identification of the optimal model parameters relative to the classification metric of concern. Ensemble models using the results of some or all of the alternative techniques were considered after all models had been established. Classification results were compared between all models with specific attention given to the performance of the alternative algorithms compared to that of logistic regression. Model assessment criteria are detailed in the following section.

Model Assessment (k-fold cross-validation)

Model assessment was done primarily through k-fold cross-validation. This process consists of splitting the data into a number of partitions or folds and then using subsets of these folds to train and test the models. In statistical modeling, the threat of over-fitting a model is well-known and occurs when a model is trained to fit data patterns too closely, resulting in the modeling of noise and anomalies in the data. Such models may be highly accurate when applied to the dataset on which they were trained, but they are often substantially less effective when applied to new data. The k-fold cross-validation technique allows for a reduced risk of overfitting, as the model is trained on different data than that which is used to test its accuracy. Periera et al. (2009) support this assessment approach by observing that if a model actually captured the relationship between a set of features and an outcome that the model would be able to perform well (i.e., classify accurately) on unseen data. The authors also note the important assumption that training and testing folds be drawn independently.

The k-folds technique is appropriate for the given context, as it allows for the final model to be trained using all of the available data (Abbott, 2014). Due to the relatively small amount of cases used in this study, k-folds is appropriate because it maximizes the use of the available information. The k-folds process randomly assigns data into k groups and then trains a model using k-1 of the groups and tests the resulting model on the single hold-out fold. This process is repeated until each of the folds has been used as the testing subset (Abbott). Abbott notes that the error rate across the k training and testing instances can serve as both a metric for accuracy as well as an index of stability, with the ideal result being similar low error rates across model building instances.

Measures of classification performance included accuracy, sensitivity, specificity, false negative rate (FNR), false positive rate (FPR), precision, false omission rate (FOR), false discovery rate (FDR), negative predictive value (NPV), positive likelihood ratio (PLR), negative likelihood ratio (NLR), and diagnostic odds ratio (DOR). As mentioned above, persistence classification studies have historically suffered from an overemphasis on the overall classification rate while neglecting to attend sufficiently to sensitivity and various other measures of error or effectiveness. In the evaluation of the results, false positive errors (cases where persisted individuals are predicted to attrite) will be more tolerable than false negative errors, as the goal is to create a model that could be practically implemented to identify students at risk of attrition. In this light, there may be more institutional tolerance for providing outreach to students who may not necessarily need it than there is to failing to provide outreach to students who would in fact benefit from additional support. Since a relatively small proportion of the overall dataset

features the attrition outcome, specificity of the results was the paramount metric for evaluation. However, high specificity alone is not sufficient, as the Type I error rate must be taken into account.

Detailed descriptive analysis are provided at the model level for each cell of the resulting classification tables. The results of these analyses were compared in order to provide a sense of the characteristics of groups that are correctly or incorrectly classified by each model. An overall case by model matrix was also examined in order to identify trends in classification across models. Attention was given to cases that were not correctly classified by any model. Specific attention was also given to whether or not there were cases that were not correctly classified by any individual model but were correctly classified by the ensemble model.

Software

Transformation, cleaning and joining of all raw data files was completed using Python 2.7. All modeling and figure creation was also completed with Python 2.7. Packages used throughout the data preparation and analysis phases include Pandas, Numpy, Scipy, Sklearn, Pylab, and Matplotlib.

CHAPTER THREE: RESULTS

Data Cleaning and Manual Feature Creation

The study initially considered the entire population of 8,135 FTFY students across six years. Several features with low variance across the entire dataset were dropped as were several categorical features which were considered irrelevant to the persistence process. The resulting dataset consisted of 115 original variables before any feature creation or transformation with 279 variables after dummy-coding and transformation. The dataset consisted of very little missing data, with missing data only present in a small number of pre-collegiate admissions figures. Due to the inability of the machine learning methods to handle missing data, 268 cases with no test scores were listwise deleted from the dataset. Notably, most of these cases were international students. The 134 cases with no available high school GPA were also listwise deleted. Additionally, the five cases where students passed away before the one-year persistence point were not considered. Finally, due to missingness on measures from the end of the first quarter, the 27 cases where students did not persist until the end of the first quarter were listwise deleted. A total of 434 cases or 5.33% of the original dataset were deleted for a final total of 7,701 cases.

The data were then split based on academic year into cross-validation and final evaluation sets. The purpose of a cross-validation set is to train models and tune parameters while also obtaining initial estimates of model generalizability. The purpose

of a held-out evaluation dataset is to test the trained and optimized model on new data that the model has not been exposed to in order to evaluate final performance and generalizability. The earliest five years of data (2011 – 2015) consisted of 6,361 cases and served as the data used to select and tune models. The final year of data (2016), consisting of 1,340 cases, was held out for final model evaluation. This process was intended to prevent data leakage and to provide a means for assessing model generalizability. The process also mimics the practical steps one would execute in order to build a model for implementation during a given year.

Several features were manually created to address scaling problems in the data. Students submit both ACT and SAT scores to fulfill admissions criteria but results from both tests are not necessarily submitted. To address this, scores from the test submitted less often, the SAT, were converted to ACT scores using available concordance tables. Scores were then merged to create a cohesive measure. Counts of incomplete, withdrawn, and failed courses were tallied and the high and low course grade values for the term were made into individual features. The proportion of courses with grades less than C- was also used as a predictor.

In order to reduce dimensionality resulting from dummy coding, the state of origin was reduced to in-state, out-of-state, and international. Dummy variables were then created for all categorical variables, including demographic features and those used to represent curriculum. Due to the number of available values for a student's academic minor, this variable was recoded to a binary feature to simply indicate whether or not the student had a declared minor by the end of the first term. Several other categorical variables, such as athlete status and Greek organization membership, were also recoded

to binary indicators. A total conduct violation variable was created to represent the sum of all conduct violations across various types during the first quarter. Three financial aid variables with excessive missingness were dropped form the analysis. A small amount of missingness was detected for variables concerning grade values, GPA, and credit hours. This missingness was determined to be due to systemic errors, so zero was imputed to represent the actual state of the value at the time of concern. The mode of 18 was imputed for the single case with a missing age value.

Cost of attendance at the institution fluctuates from year to year, with a 17.45% increase from the earliest to the most recent cohort included in the analysis. In order to account for the proportion of total cost covered by each award, the magnitudes of the financial aid figures were adjusted for each year by a constant in order to increase their comparability.

All cases were randomly shuffled prior to analysis to prevent any possible bias from the k-folds split process. All remaining variables were z-score normalized or min-max transformed during each train-test iteration of cross validation. MLP and Gaussian naïve Bayes models were balanced using SMOTE within every train-test iteration. SMOTE was not used to balance logistic regression data during cross-validation. Logistic regression balancing was achieved through a class weight parameter.

Descriptive Analysis

Persistence and attrition rates as well as cohort size can be seen in Table 1 below. As the table indicates, there was, on average, over a recent six-year period, a loss of 13.82% of the FTFY class each year in the retained data.

Table 5

One-Year Persistence and Attrition Rates by Cohort

Cohort	N	Attrited	Persisted
201170WK3N	1,140	13.16%	86.84%
201270WK3N	1,131	13.62%	86.38%
201370WK3N	1,365	13.41%	86.59%
201470WK3N	1,359	12.80%	87.20%
201570WK3N	1,366	13.54%	86.46%
201670WK3N	1,340	13.21%	86.79%
Total	7,701	13.28%	86.72%

Note. Cohort = label for each of the six cohorts of FTFY students; N = number of students in cohort population; Attrited = proportion of students that did not persist to week three of the second year; Persisted = proportion of students that were enrolled at week three of the second year

Variable-level descriptive statistics, disaggregated by cross-validation and evaluation sets, for a selection of demographic, academic, and social features can be viewed in Appendix C. As suggested by the consistency of these figures, there was a great deal of similarity between average student attributes in the two datasets.

Cross Validation

Parameter tuning and optimization for each model was facilitated by a grid search process in which 10-fold cross-validation was conducted using all specified permutations of parameter settings. Models were then selected during that process according to the highest score on the specified evaluation metric, F-Beta, which is denoted by the following equation:

$$F_\beta = (1+\beta^2) \times ((\text{precision} \times \text{recall}) / (\beta^2 \times \text{precision}) + \text{recall})) \quad (4)$$

The grid search cross-validation process was then repeated again this time selecting the model optimized for F1, which is specified below.

$$F_1 = 2 \times ((\text{precision} \times \text{recall}) / (\text{precision} + \text{recall})) \quad (5)$$

Conducting a cross-validated grid search optimized for these metrics was done in order to allow greater latitude for model selection. Model sensitivity is a primary concern of the research and F-Beta provides greater weight to sensitivity in its computation. This can, however, result in models with lower than desired precision. F1 computes an unweighted harmonic mean of sensitivity and precision, providing an alternative to models optimized for F-Beta that may have undesirably low precision.

Logistic Regression

Examined regularization strengths of 0.00001, 0.001, 0.1, 10.0 and 1,000.0 were evaluated. The k-best predictors were also considered with k evaluated at 25, 50, 75 and 100. Balanced and unbalanced class weighting was also taken into account. The highest scoring logistic regression model, optimized for F-Beta, utilized 75 predictors, balanced class weighting, inverse regularization strength of 0.1, and the LIBLINEAR linear classifier. The identical logistic regression model also performed best when optimizing for F1. Means and standard deviations for scoring metrics of concern for the highest performing cross-validated test set models are displayed in Table 5. Final models were retrained on all training and testing data using the parameters from the cross-validated model with the highest F-Beta score. These models were then used to predict the

outcome of the held-out evaluation data. Table 2 displays the confusion matrix for the final prediction on the evaluation dataset, with true positive, true negative, false positive, and false negative rates noted in parentheses. A confusion matrix is a two by two matrix where the sum of the values in the first row is equal to the actual number of negative cases in the data and where the sum of the second row is equal to the actual number of positive cases in the data. The sum of the first column in a confusion matrix represents the total number of negative case predictions, while the sum of the second column represents the sum of positive case predictions. The cells of the matrix represent the intersection of predicted and actual classes. Table 6 displays the hold-out evaluation metrics for these predictions. Note that all cross-validation test metrics and final evaluation results are identical for the F-Beta and F1 optimized logistic regression models, as the models are identical. Plots of precision and recall as a function of decision threshold as well as ROC curves can be found in Appendix D.

Table 6

Logistic Regression Final Evaluation Confusion Matrix

	Predicted			
	F-Beta		F1	
	Persist	Attrit	Persist	Attrit
Persist	890 (76.53%)	273 (23.47%)	890 (76.53%)	273 (23.47%)
Attrit	69 (38.98%)	108 (61.02%)	69 (38.98%)	108 (61.02%)

Note. Predicted = model predictions. Sum of columns represent predicted cases for persist and attrit outcomes; F-Beta = model optimized for F-Beta; F1 = model optimized for F1; Persist = the number of students who actually persisted or were predicted to persist; Attrit = the number of students who actually attrited or were predicted to attrit

Multilayer Perceptron Neural Network

The MLP utilized the grid search strategy, examining single hidden layer networks with 25 and 50 neurons as well as two hidden layer networks with 25 and 50 neurons in each layer. The alpha regularization parameter was evaluated at 0.0001, 0.001 and 0.01. As with the logistic regression model, the k best predictors were used, with k evaluated at 50, 75, 100, 125 and 150. Both the logistic and hyperbolic tangent activation functions were examined. All permutations of the potential model hyperparameters were examined. An adaptive learning rate and maximum of 1,000 epochs were used for every MLP. Due to MLP input requirements, all data were transformed using a min-max scaler in order to bind values between 0 and 1, versus the z-score normalization process utilized for the other model types. The cross-validation data were balanced using SMOTE. The MLP model, optimized for F-Beta, utilized 75 predictors, a single hidden layer with 50 neurons, a regularization parameter of 0.01 and a logistic activation function. The MLP model, optimized for F1, utilized 75 predictors, two hidden layers with 25 neurons each, a regularization parameter of 0.001 and a logistic activation function. Table 3 displays the confusion matrix for the evaluation of both MLP models on the hold-out set. Cross-validation test metrics can be seen, along with those metrics for all model types, in Table 5. Final metrics for hold-out set performance are displayed in Table 6.

Note that the cross-validation metrics are very similar for both MLPs, with slightly higher recall for the F-Beta-optimized model and slightly higher precision and overall accuracy for the F1-optimized model. Both models scored identically on F-Beta and ROC AUC metrics. The similarity of these models is further evidenced by the confusion matrices resulting from their application to the hold-out set. In this regard, the

F-Beta model slightly outperforms the F1 model with better ability to identify the persisted students. Precision and overall accuracy are higher for the F-Beta MLP on the hold-out data, although these metrics are higher for the F1 MLP during training and testing. Plots of precision and recall as a function of decision threshold as well as ROC curves for the MLP models can be found in Appendix D.

Table 7

Multilayer Perceptron Final Evaluation Confusion Matrix

	Predicted			
	F-Beta		F1	
	Persist	Attrit	Persist	Attrit
Persist	831 (71.45%)	332 (28.55%)	825 (70.94%)	338 (29.06%)
Attrit	67 (37.85%)	110 (62.15%)	67 (37.85%)	110 (62.15%)

Note. Predicted = model predictions. Sum of columns represent predicted cases for persist and attrit outcomes; F-Beta = model optimized for F-Beta; F1 = model optimized for F1; Persist = the number of students who actually persisted or were predicted to persist; Attrit = the number of students who actually attrited or were predicted to attrit

Naïve Bayes

The Gaussian naïve Bayes (GNB) model, optimized for F-Beta, takes no special parameters, therefore the only parameter tuning concerned the k-best predictors. K was evaluated at 25, 50, 75, 100, 125 and 150. The final model, optimized for F-Beta, utilized 100 predictors and a balanced cross-validation dataset created by the application of SMOTE. The naïve Bayes model with F1 optimization, also balanced with SMOTE, utilized 50 predictors. Confusion matrices for the GNB models are displayed in Table 4.

Average cross-validation test metrics are shown in Table 5, while hold-out evaluation metrics are displayed in Table 6. Plots of precision and recall for the GNB models as well as ROC curves can be found in Appendix D.

Table 8

Naïve Bayes Final Evaluation Confusion Matrix

	Predicted			
	F-Beta		F1	
	Persist	Attrit	Persist	Attrit
Persist	627 (53.91%)	536 (46.09%)	919 (79.02%)	244 (20.98%)
Attrit	57 (32.20%)	120 (67.80%)	87 (49.15%)	90 (50.85%)

Note. Predicted = model predictions. Sum of columns represent predicted cases for persist and attrit outcomes; F-Beta = model optimized for F-Beta; F1 = model optimized for F1; Persist = the number of students who actually persisted or were predicted to persist; Attrit = the number of students who actually attrited or were predicted to attrit

Note that the optimization metric (F-Beta or F1) for logistic regression and MLP made little or no difference in terms of model classification abilities. However, for GNB this process resulted in large variations in both cross-validation and hold-out results. The F-Beta-optimized GNB model demonstrated much higher recall in cross-validation (M = 0.74, SD = 0.08 vs M = 0.51, SD = 0.11) at the cost of lower precision (M = 0.16, SD = 0.01 vs M = 0.25, SD = 0.04). These characteristics were largely generalizable to the hold-out data. Both GNB models demonstrated desirable characteristics. The F-Beta recall score was very high compared to any of the tested models, while its precision was among the lowest. The F1-optimzed GNB model, on the other hand, makes more conservative predictions, boasting a medium recall score, and a higher precision score.

Table 5

Cross-Validation Test Set Metrics

Metric	Model	F-Beta *M(SD)*	Recall *M(SD)*	Precision *M(SD)*	ROC AUC *M(SD)*	F1 *M(SD)*	Accuracy *M(SD)*
F-Beta	LR	0.48 (0.04)	0.63 (0.05)	0.25 (0.02)	0.72 (0.03)	0.35 (0.03)	0.70 (0.02)
	MLP	0.47 (0.03)	0.63 (0.05)	0.23 (0.02)	0.71 (0.03)	0.34 (0.02)	0.67 (0.03)
	GNB	0.43 (0.02)	0.74 (0.08)	0.16 (0.01)	0.65 (0.03)	0.26 (0.01)	0.45 (0.07)
F1	LR	0.48 (0.04)	0.63 (0.05)	0.25 (0.02)	0.72 (0.03)	0.35 (0.03)	0.70 (0.02)
	MLP	0.47 (0.03)	0.61 (0.04)	0.24 (0.02)	0.71 (0.03)	0.34 (0.02)	0.69 (0.02)
	GNB	0.41 (0.03)	0.51 (.11)	0.25 (0.04)	0.69 (0.02)	0.33 (0.03)	0.71 (0.10)

Note: F-Beta = weighted harmonic mean of precision and recall with β set to 2; Recall = TP / (TP+FN); Precision = TP / (TP+FP); ROC_AUC = area under the receiver operating characteristic curve; F1 = harmonic mean of precision and recall; Accuracy = (TP + TN) / (TP +TN+FP+FN); M(SD) = mean and standard deviation of 10-fold cross-validation testing; Metric = evaluation metric for which the model has been optimized; Model = model type; LR = logistic regression; MLP = multilayer perceptron; GNB = Gaussian naïve Bayes

Table 6

Final Evaluation Set Metrics

Metric	Model	F-Beta	Recall	Precision	ROC AUC	F1	Accuracy
F-Beta	LR	0.50	0.61	0.28	0.69	0.39	0.74
	MLP	0.48	0.62	0.25	0.67	0.36	0.70
	GNB	0.44	0.68	0.18	0.61	0.29	0.56
F1	LR	0.50	0.61	0.28	0.69	0.39	0.74
	MLP	0.48	0.62	0.25	0.67	0.35	0.70
	GNB	0.44	0.51	0.27	0.65	0.35	0.75

Note: F-Beta = weighted harmonic mean of precision and recall with β set to 2; Recall = TP / (TP+FN); Precision = TP / (TP+FP); ROC_AUC = area under the receiver operating characteristic curve; F1 = harmonic mean of precision and recall; Accuracy = (TP + TN) / (TP +TN+FP+FN); Metric = evaluation metric for which the model has been optimized; Model = model type; LR = logistic regression; MLP = multilayer perceptron; GNB = Gaussian naïve Bayes

Research Questions

Research question 1a. Which of the examined models results in the most desirable classification of students into attrition and persistence groups?

Based on the metrics used to optimize the models, F-Beta and F1, the logistic regression model outperformed all other models on the hold-out set with scores of 0.50 and 0.39, respectively. The F-Beta-optimized MLP demonstrated the next highest scores for these metrics on the hold-out set with F-Beta = 0.48 and F1 = 0.36. The logistic regression ROC AUC score of 0.69 and precision of 0.28 were also higher than any of the other final models. Both MLPs achieved recall scores of 0.62, which is slightly higher than the recall score of 0.61 achieved by the logistic regression model. The F-Beta-optimized GNB model achieved the highest recall on the hold-out set with a score of

0.68. However, when considered alongside their lower precision scores (0.25 for both MLPs and 0.18 for the F-Beta-optimized GNB), the MLP and GNB model performances are diminished in regard to a well-balanced performance on recall and precision. Notably, the F1-opimized GNB model resulted in the highest overall accuracy, with an accuracy score of 0.75 compared to the 0.74 accuracy score achieved by the logistic regression model. In regard to the classification of most concern, the attriters, the GNB model was less sensitive (recall = 0.51) although fairly precise (precision = 0.27). The F1-optimized GNB did, however, achieve a higher correct classification rate for persisting students than any other model with a true negative rate of 79.02% on the hold-out set. The logistic regression model achieved a true negative rate of 76.53% during the final evaluation, while the highest performing MLP in this regard, the F-Beta-optimized model, achieved a true negative rate of 71.54%.

Further evidence of the relative superiority of the logistic regression model in this context can be seen in the plot of all final model ROC curves seen in Figure 2. The plot demonstrates that the logistic regression model achieved a higher recall to false positive rate across almost any decision boundary point. Note that there is much consistency in the general shape of the curves, with differences being mostly due to height of the curves and the area beneath them. The logistic regression model and the MLP optimized for F-Beta were very similar in their tradeoff between recall and FPR along lower decision thresholds, although the two diverge at approximately 0.15 FPR. The 0.15 FPR point is where the steepness of most of the curves begins to decline and to take on a more gradual trajectory. Notably, the logistic regression model ROC curve was most pronounced in terms of its divergence from the other models at approximately the 0.35 FPR point, where

it was able to achieve recall of approximately 75.00%. As the curves demonstrate, the final logistic model resulted in the highest AUC at 0.69. Both MLPs achieved AUC scores of 0.67. The proximity of AUC scores between the MLPs and the logistic regression model can be seen in Figure 2, and it is notable to highlight that there are multiple decision thresholds at which the MLPs and the logistic regression are able to achieve nearly identical recall to FPR ratios.

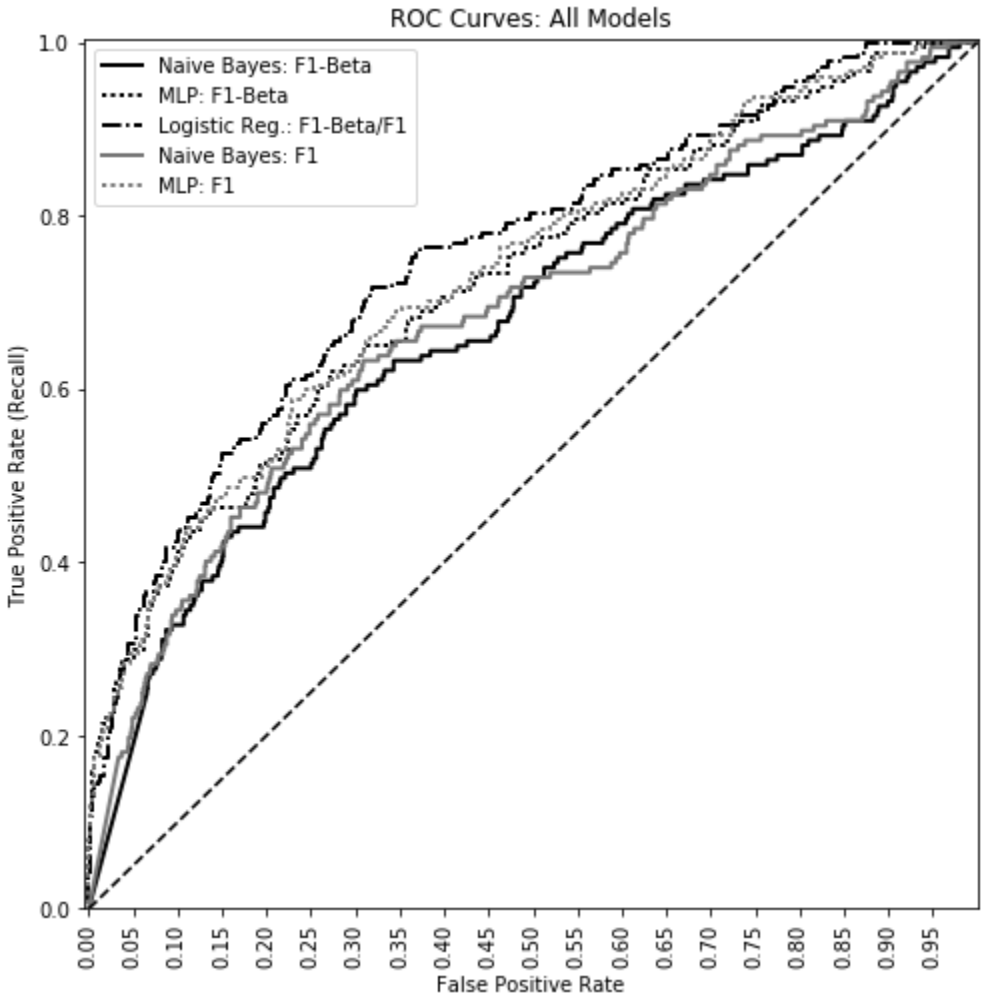


Figure 10. Final Evaluation ROC Curve

Research question 1b. Does the ensemble model surpass the best performing individual model in terms of the examined classification metrics?

The first ensemble classifier to be tested was a simple majority rule voting classifier (MVE) in which cases were assigned to the class predicted by the majority of models included in the ensemble. This classifier required no weight optimization through cross-validation since the weights were established *a priori*. Table 7 displays the confusion matrix for the best individual performing model (logistic regression) and the majority voting ensemble. Final performance evaluation metrics are displayed in Table 8.

Table 7

Majority Voting Ensemble vs Logistic Regression Confusion Matrix

	Predicted			
	Majority Voting Ensemble		LR	
	Persist	Attrit	Persist	Attrit
Persist	833 (71.63%)	330 (28.37%)	890 (76.53%)	273 (23.47%)
Attrit	65 (36.72%)	112 (63.28%)	69 (38.98%)	108 (61.02%)

Note. Predicted = model predictions. Sum of columns represent predicted cases for persist and attrit outcomes; Majority Voting Ensemble = represents MVE model scores; LR = individual logistic regression model scores; Persist = the number of students who actually persisted or were predicted to persist; Attrit = the number of students who actually attrited or were predicted to attrit

Table 8

Final Evaluation Ensemble Set Performance vs Logistic Regression

Model	F-Beta	Recall	Precision	ROC AUC	F1	Accuracy
LR	0.50	0.61	0.28	0.69	0.39	0.74
MVE	0.49	0.63	0.25	0.67	0.36	0.71
EM2-FB	0.49	0.62	0.27	0.68	0.37	0.72
EM2-F1	0.49	0.61	0.27	0.68	0.37	0.73

Note: F-Beta = weighted harmonic mean of precision and recall with β set to 2; Recall = TP / (TP+FN); Precision = TP / (TP+FP); ROC_AUC = area under the receiver operating characteristic curve; F1 = harmonic mean of precision and recall; Accuracy = (TP + TN) / (TP +TN+FP+FN); Metric = evaluation metric for which the model has been optimized; Model = model type; LR = logistic regression; MLP = multilayer perceptron; GNB = Gaussian naïve Bayes

As Tables 8 and 9 demonstrate, the majority voting ensemble assigning equal weight to each classifier resulted in four additional correct classifications of attrition above the logistic regression model, but these were at the cost of 57 additional false attrition classifications. Additionally, the MVE classified 57 fewer cases of persistence correctly. All other metrics of concern other than recall were lower for the MVE than for the best performing individual model, logistic regression. The evaluative metrics of greatest concern, F-Beta and F1, were 0.48 and 0.35 for the MVE on the hold-out set while the logistic regression model achieved final evaluation scores of 0.50 and 0.39 on these metrics, respectively. While the MVE resulted in slightly higher recall, the overall performance compared to the logistic regression model was diminished when using majority rule voting with equal weighting for all classifiers.

In addition to the equally weighted majority voting classifier, several other permutations of classifiers and weights were examined. In order to assess the feasibility

of alternative ensemble models utilizing k classifiers, all weight permutations ranging from 0 to k-1 were examined through a 10-fold cross-validation process. Ensemble model weights were optimized for F-Beta and F1. The inclusion of the 0 weight coefficient allowed for the initial detection of a classifier that could potentially outperform the logistic regression model on the hold-out evaluation data. Through this process, any identified weighting scheme where weight was assigned to a classifier other than the logistic regression model indicated enhanced performance through an ensemble approach and implied a potential for these improvements to generalize to the hold-out set. By evaluating the cross-validation results using all of the specified weighting permutations, including the permutation with all but the logistic regression weight set to 0, this process allowed for an empirical selection of ensemble model weights based on F-Beta and F1 results.

Cross-validation was not used for the MVE model since weights and classifiers were established *a priori.* However, in order to make an unbiased selection of classifiers and associated weights, these decisions must be made based on cross-validation performance and not on hold-out set evaluation scores from random combinations of classifiers and weights. For this reason, the additional ensemble models were created through a weighting and classifier selection optimization process within cross-validation folds. Final models were then evaluated against logistic regression on the hold-out set.

The second ensemble model considered four classifiers, the logistic regression model, the MLP optimized for F-Beta and both the F1-optimized and F-Beta-optimized GNB models. The F1-optimized MLP model was not considered due to its similarity to the F-Beta-optimized MLP model and also for purposes of reducing the size of the

optimal weight search space. Weights for the second ensemble were first optimized for F-Beta. The results indicated that the optimal weighting scheme within the cross-validation set was weight = 2 for the F-Beta-optimized MLP and weight = 3 for the logistic regression. Ensemble models and associated weights are displayed in Table 9 along with F-Beta and F1 cross-validation scores. As the table indicates, optimal F1 weighting for the second ensemble model assigned weight = 1 to the F-Beta-optimized MLP and weight = 3 to the logistic regression. Notably, within cross-validation, the optimal weighting for both F1 and F-Beta was achieved with input from the MLP. When optimizing for F-Beta, the solution which assigned all weight to the logistic regression alone scored the sixth highest of all other solutions, while the logistic regression alone scored third highest when optimizing for F1. However, differences between the top scoring cross-validation weighting solutions were minimal. No GNB models were selected for a role in this ensemble. Cross-validation results and associated weighting for the examined ensemble combinations are displayed in Table 9.

Table 9

Cross-Validation Scores and Weights for Empirically-Derived Ensembles

	MLP F-Beta	LR	GNB F-Beta	GNB F1	F-Beta	F1
Model	Weight				*M(SD)*	
EM2: F-Beta	2	3	0	0	0.48 (0.04)	0.36 (0.03)
EM2: F1	1	3	0	0	0.48 (0.03)	0.36 (0.02)

Note. MLP F-Beta = MLP model optimized for F-Beta; LR = individual logistic regression model; GNB F-Beta = GNB model optimized for F-Beta; GNB F1 = GNB model optimized for F1; F-Beta = F-Beta score; F1 = F1 score; Model = name of empirically weighted ensemble; Weight = weight assigned to individual models in each empirically weighted ensemble; M(SD) = mean and standard deviation for F-Beta and F1 scores

The improved scores for F-Beta and F1 observed during cross-validation for the second ensemble model suggested a potentially improved classifier for use on the hold-out evaluation data. However, as Table 8 shows, the final F-Beta and F1 evaluation metrics for both iterations of the second ensemble were lower than those obtained by the logistic regression model alone. The F-Beta-optimized iteration of the second ensemble, however, did achieve a slightly higher recall than the logistic regression model (0.62 vs 0.61). Both iterations of the second ensemble achieved F-Beta scores of 0.49 and F-1 scores of 0.37, compared to 0.50 and 0.39, respectively, for the logistic regression model alone. Confusion matrices for both versions of the second ensemble compared to the logistic regression model are presented in Table 10, which demonstrates objectively diminished performance by the second ensemble optimized for F1 and an improvement of only one correct attrition classification by the F-Beta optimized model at a cost of a higher false positive rate and diminished ability to correctly classify cases of persistence.

Table 10

Confusion Matrices for Ensemble Model Two, Compared to Logistic Regression

	EM2 F-Beta-Optimized		EM2 F1-Optimized		LR	
	Persist	Attrit	Persist	Attrit	Persist	Attrit
Persist	862	301	871	292	890	273
	(74.12%)	(25.88%)	(74.89%)	(25.11%)	(76.53%)	(23.47%)
Attrit	68	109	69	108	69	108
	(38.42%)	(61.58%)	(38.98%)	(61.02%)	(38.98%)	(61.02%)

Note. EM2 F-Beta-Optimized = empirically derived model optimized for F-Beta; EM2 F1-Optimized = empirically derived model optimized for F1; LR = individual logistic regression model; Persist = actual or predicted number of persisted students; Attrit = actual or predicted number of attrited students; parenthetical figures represent true positive rate, false positive rate, false negative rate and true negative rate

Notably, both iterations of the second ensemble demonstrated excellent generalizability on the hold-out set with improvements from the average F-Beta and F1 scores obtained through cross-validation. Improvement in scores from cross-validation to hold-out evaluation were seen for both MLPs during individual model building and evaluation. However, these improvements from initial model assessment on cross-validation data to final evaluation on the held-out data were also seen, but to a larger degree, in the case of the individual logistic regression model. Overall, while some ensemble models showed promise during cross-validation, none were able to outperform the individual logistic regression model in terms of F-Beta and F1 scores on the held-out evaluation set. In all examined ensemble models the integration of predictions from any other model diminished the performance of the logistic regression model alone.

Research question 1c. Do any of the alternative algorithms perform better than the logistic regression model in regard to the metrics of concern?

While both MLP models and the F-Beta-optimized GNB model all accurately identified more cases of attrition in the hold-out dataset than the logistic regression model (110 and 120, respectively, versus 108), the logistic regression model, as detailed under research question 2a, still outperformed any other individual model on the evaluative metrics of concern. Additional evidence of both the distinction and similarity between the models can be seen below in Figures 3, 4, and 5, which show the precision-recall curves for the five optimized models. While both MLPs have a jagged and slightly more gradual descent in precision than the logistic regression model, which drops off sharply in precision around 0.12 to 0.13 recall, the logistic regression model was able to maintain a slightly higher and somewhat smoother descent in precision across the 0.40 to 0.60 recall range. The logistic regression model's ability to maintain this slightly higher level of precision across this range results in its superior hold-out evaluation performance relative to the MLPs. However, the models clearly exhibit a similar tradeoff between these two metrics.

The GNB models, as Figure 5 demonstrates, have a gradually descending precision level across most levels of recall which resembles the curves at those points for the MLP and logistic regression models. However, the F1-optimized GNB was unable to attain precision higher than approximately 0.44 at any level of recall, while the F-Beta-optimized GNB was unable to attain precision higher than approximately 0.38. Both GNB models had a relatively low ceiling for precision even at low levels of recall, however, when optimized for F1 and F-Beta, the GNB model was able to produce

comparable, although inferior, results to the MLPs and logistic regression models. The additional predictive value of the GNBs, in regard to ability to correctly identify cases not detected by the MLPs and logistic regression models, is discussed below in regard to research question 3a.

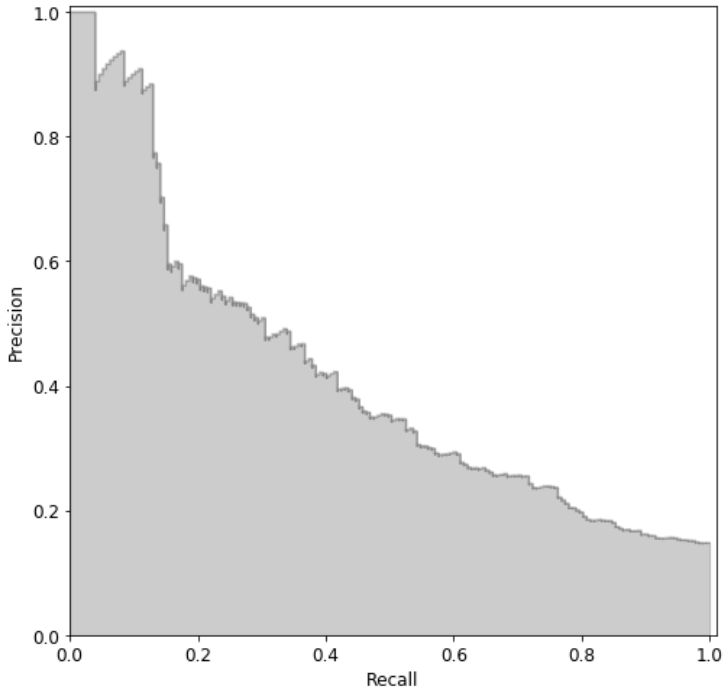


Figure 11. Precision-Recall Curve: Logistic Regression

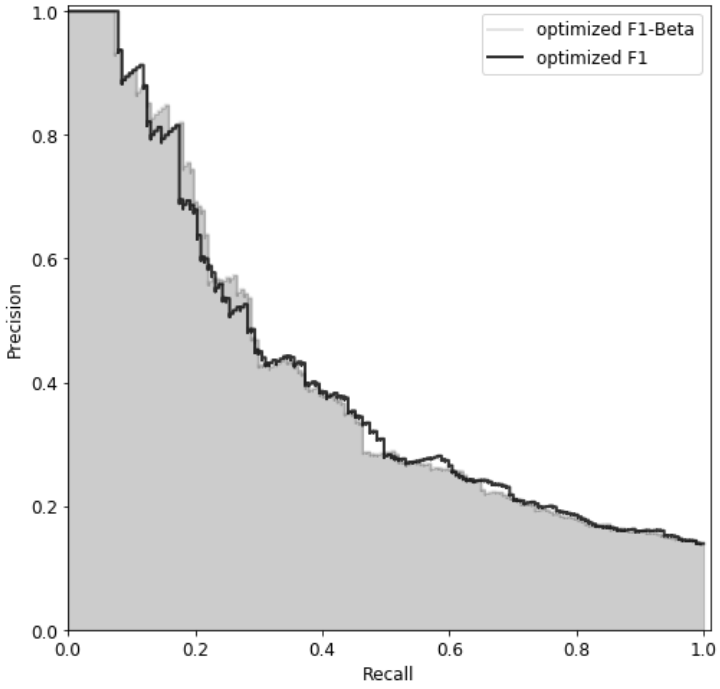


Figure 12. Precision-Recall Curve: MLP optimized for F-Beta and F-1

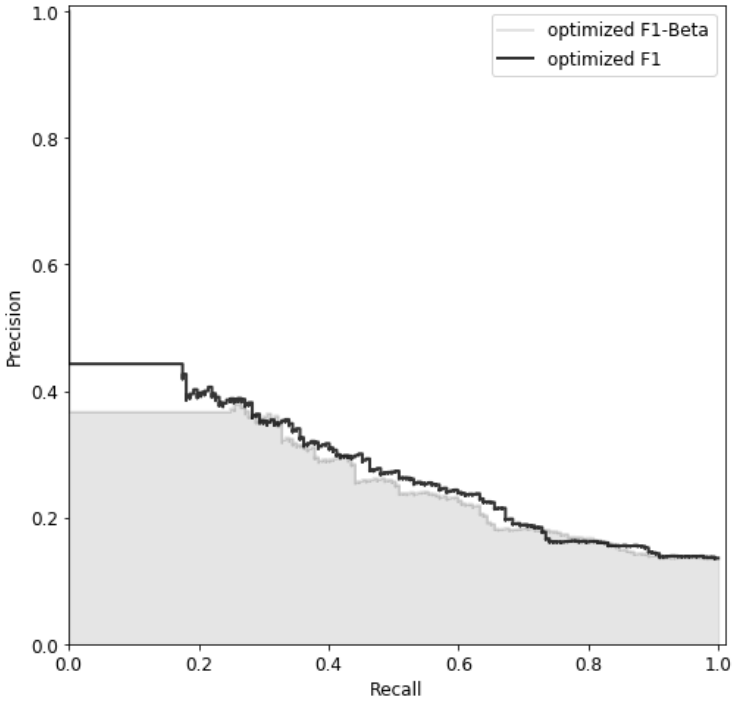


Figure 13. Precision-Recall Curve: Naïve Bayes

Research question 2a. Which features or composite features result in the best performing classification model?

The implementation of a univariate feature selection method, select K best, combined with a grid search process evaluating multiple values of K resulted in three different utilized sets of variables. The logistic regression model and both MLPs utilized 75 of the available features, while the F-Beta-optimized GNB model utilized 100 features and the F1-optimized GNB utilized 50 features. See Appendix F for a list of the top 50, top 75, and top 100 predictors, selected through a univariate process.

As noted above, the best performing model, logistic regression, utilized 75 features. It should be noted that the feature selection process for each of the final models was implemented using all of the cross-validation data, so feature selection was not performed on the held-out evaluation data directly, as univariate feature selection requires a known outcome with which to correlate individual features. The importance of individual features for the logistic regression model was evaluated by rank ordering the absolute values of the products of model coefficients and standard deviations for the relevant variables. This method was implemented in order to reduce over-inflation of dummy coded variable importance in the feature ranking. Since variables were standardized prior to variable selection, the majority of standard deviations are 1.00. However, standard deviations for dummy coded categorical variables vary. Table 11 displays the top 20 predictive features for the final logistic regression model. For a complete list of ranked features see Appendix G.

Table 11

Top 20 Predictive Features for Logistic Regression

Feature	Coefficient	Odds	SD	Coeficient x SD	ABS
FUND_SOURCE_DESC_Undergraduate Discount	-0.57	0.56	1	-0.57	0.57
TOTAL_ACCEPT_AMOUNT	-0.54	0.58	1	-0.54	0.54
ADMINISTRATOR_RATING	-0.27	0.76	1	-0.27	0.27
FUND_SOURCE_DESC_State Funding	0.25	1.29	1	0.25	0.25
mean_GradeVal_Q1	-0.21	0.81	1	-0.21	0.21
GIFT_OR_SELF_HELP_Gift Aid	-0.19	0.83	1	-0.19	0.19
CREDITS_EARNED_Q1	-0.19	0.83	1	-0.19	0.19
CREDITS_ATTEMPTED_Q1	-0.18	0.83	1	-0.18	0.18
Total_Conduct_Q1	-0.18	0.84	1	-0.18	0.18
TOTAL_CREDITS_Q1	0.18	1.19	1	0.18	0.18
TOTAL_OFFER_AMOUNT	-0.17	0.84	1	-0.17	0.17
HoldsNew_Q1	-0.16	0.85	1	-0.16	0.16
min_GradeVal_Q1	-0.16	0.85	1	-0.16	0.16
Activity_Count_Q1	-0.14	0.87	1	-0.14	0.14
FIN_AID_TYPE_Scholarship	0.14	1.15	1	0.14	0.14
Proportion_belowCminus_Q1	0.13	1.14	1	0.13	0.13
FUND_SOURCE_DESC_Departmental Funded Schl	0.13	1.14	1	0.13	0.13
Cares_Submissions_Q1	0.13	1.13	1	0.13	0.13
W_Count_Q1	-0.12	0.89	1	-0.12	0.12
Sexual_EEO_Q1	0.12	1.12	1	0.12	0.12
Endangerment_Weapons_Provoke_Q1	0.11	1.12	1	0.11	0.11

Note: Feature = variable name from dataset; Coefficient = regression coefficient from final model; Odds = the odds associated with the coefficient (the exponentiation of the coefficient); SD = the standard deviation

for the feature in the held-out evaluation dataset; Coefficient X SD = the product of a feature's regression coefficient and its standard deviation; ABS = the absolute value of the Coefficient X SD value and the value on which the features are ordered in the table.

As Table 11 demonstrates, four of the top ten predictors were related to financial aid (FUND_SOURCE_DESC_Undergraduate Discount, TOTAL_ACCEPT_AMOUNT, FUND_SOURCE_DESC_State Funding, GIFT_OR_SELF_HELP_Gift Aid). Four of the top ten predictors were also related to first quarter academic performance (mean_GradeVal_Q1, CREDITS_EARNED_Q1, CREDITS_ATTEMPTED_Q1, TOTAL_CREDITS_Q1). One of the top ten predictors was related to pre-collegiate qualifications (ADMINISTRATOR_RATING), and one was related to student conduct (Total_Conduct_Q1). For interpretation of model coefficients, note that the endogenous persistence variable was coded as 0 for persistence and 1 for attrition. All but two coefficients for the top ten predictors were negative, suggesting that increases in those areas are associated with greater likelihood of persistence. The coefficients for FUND_SOURCE_DESC_State Funding and TOTAL_CREDITS_Q1 were positive, suggesting that higher values for those variables are associated with greater likelihood of attrition. Interestingly, the Total_Conduct_Q1 variable, which represents the total number of student conduct violations for the term, had a negative coefficient, suggesting that higher conduct violations were associated with increased likelihood of persistence. Implications of variable importance and coefficients are discussed in more depth in the next chapter.

The odds associated with each feature, as seen in Table 11, express the percent increase or decrease in the odds of being in the attrition group associated with one standard deviation increase on the given feature, holding all other variables constant. All

variables with odds greater than 1.00 can be interpreted as having greater quantities of that variable associated with increased odds for being in the attrition class, while all features with odds less than 1.00 can be interpreted as having greater values associated with decreased odds of being classified as attrition. As Table 11 demonstrates, a one standard deviation increase of FUND_SOURCE_DESC_Undergraduate Discount is associated with an approximately 44.00% reduction of the odds of being classified as an attriter, while a single standard deviation increase in TOTAL_ACCEPT_AMOUNT is associated with an approximately 42.00% reduction in the odds of an attrition classification. The two variables in the top ten predictors with positive coefficients were FUND_SOURCE_DESC_State Funding and TOTAL_CREDITS_Q1. The odds associated with these features were 1.29 and 1.19, respectively, indicating an approximate 29.00% and 19.00% increase in the odds of being classified as an attriter associated with a single standard deviation increase in the variables, holding all else constant.

Of the features from the final logistic regression model ranking in importance from 11 to 20, three were related to student conduct violations (Cares_Submissions_Q1, Sexual_EEO Q1, Endangerment_Weapons_Provoke_Q1), three were related to academic performance (min_GradeVal_Q1, Proportion belowCminus Q1, W_Count_Q1), two were related to financial aid (FIN_AID_TYPE_Scholarship, FUND_SOURCE_DESC_Departmental Funded Schl), one was related to student accounts (HoldsNew_Q1), and one was related to student co-curricular involvement (Activity_Count_Q1). As can be seen from Table 11 demonstrates, the proportional change in the odds given one standard deviation of change in the values while holding all

else constant for these predictors ranged between positive or negative 0.15 to 0.11, a smaller magnitude than seen with the top ten predictors.

Research question 2b. Which of the rarely explored data elements are the most powerful predictors?

The top 20 predictors from the logistic regression model consisted of both conventional and rarely-explored predictors. The total financial aid award (TOTAL_ACCEPT_AMOUNT) emerged as the second best predictor based on the ranking method. However, awards broken out by fund category (FUND_SOURCE_DESC_Undergraduate Discount, FUND_SOURCE_DESC_State Funding, FIN_AID_TYPE_Scholarship, FUND_SOURCE_DESC_Departmental Funded Schl) also ranked in the top 20 predictors, with the undergraduate discount rate emerging as the best overall predictor. Student account data, such as various types of holds and aggregations of those holds, is also considered as rarely explored information. Two of the hold variables (HoldsNew_Q1, Active_Holds_Q1) emerged in the top 30 predictors. Interestingly, the odds of 0.85 for HoldsNew_Q1 suggested a decrease in the odds of being in the attrition group associated with greater numbers of new holds while the odds of 1.10 for Active_Holds_Q1 suggested an approximately 10.00% increase in the odds of being classified as an attriter for each standard deviation increase in active holds.

The mean_GradeVal_Q1 feature emerged as the most powerful indicator related to academic performance. While this feature used common academic performance data in its creation, it is a computed field meant to be a potential alternative to the grade point average field which is a weighted calculation of course grade values. The feature, Proportion below Cminus Q1, another computed feature which represents the proportion

of grades received in the first term which were less than a C-, was also found to be one of the 20 best predictors.

The specific campus housing building in which students reside is also considered to be rarely explored information. Dummy coded features representing two different housing locations (BUILDING_DESC_Q1_McFarlane Hall, BUILDING_DESC_Q1_Centennial Towers) also emerged in the top 40 predictors, with residence in the former building decreasing the odds of attrition and residence in the latter building increasing the odds of attrition.

Some of the most powerful predictors in the non-traditional data were the conduct violations features, with eight of these variables (Total_Conduct_Q1, Cares_Submissions_Q1, Sexual EEO Q1, Endangerment_Weapons_Provoke_Q1, Academic_Difficulty_Q1, Dishonesty_Q1, Drugs_Alcohol_Q1, Mental_Heallth_Q1) appearing in the top 40 predictors for the final logistic regression model. Notably, with the exception of Total_Conduct_Q1 and Drugs_Alcohol_Q1, all student conduct predictors had negative coefficients, suggesting that greater numbers of violations or cases are associated with increased odds of attrition. Total_Conduct_Q1 represents the total number of conduct violations during the quarter, the majority of which, on average, are related to drug or alcohol possession and use. The feature, Drugs_Alcohol_Q1, represents the total number of drug or alcohol possession cases during the quarter. The observation that increased incidence of total conduct and drug or alcohol violations was associated with reduced odds of attrition will be discussed further in the following chapter.

Research question 3a. To what extent do the models differ in terms of the cases that they accurately classify or misclassify?

When applied to the hold-out set, complete correct consensus was achieved on 68 (38.43%) of the 177 attrition cases. Accurate attrition consensus was reached by four or more models on 91 (51.41%) of the cases, while three or more models correctly predicted 113 (63.84%) of the attrition cases. Accurate attrition consensus was reached for 125 (70.62%) of the cases. There were only 16 records (9.04% of attrition cases) where only one model correctly classified an instance of attrition, with one of those detected by the F1-optimized MLP, three detected by the logistic regression model, nine detected by the GNB model optimized for F-Beta and three detected by the F1-optimized GNB model. There were 36 (20.34%) attrition cases in the hold-out dataset where no models made an accurate prediction. Figure 6 provides a visual representation of the case classifications.

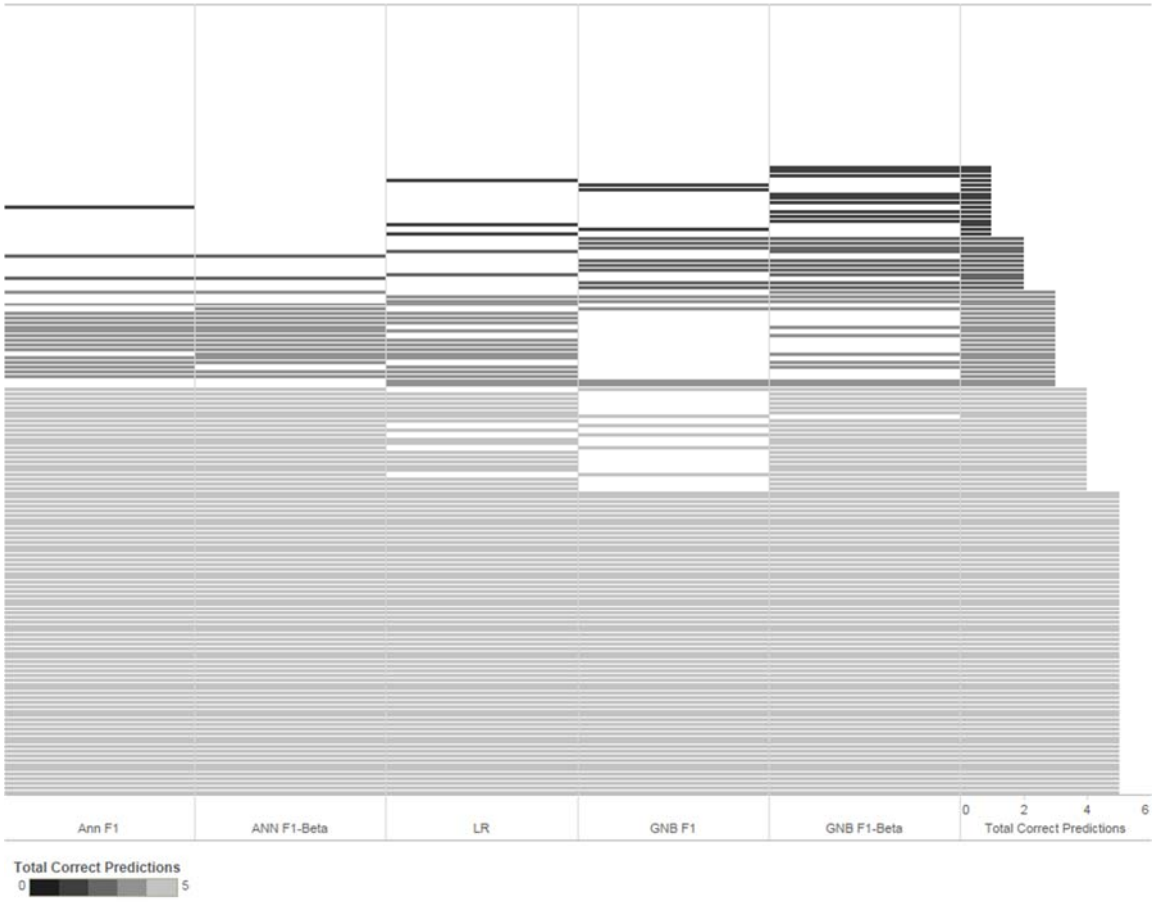


Figure 14. Correctly Classified Attrition Cases from Hold-Out by Model Type by Consensus

The y-axis of the figure represents the 177 attrition cases. The x-axis is comprised of the five different models. Shaded bars represent instances where the corresponding classifier accurately predicted a given case. The darkest bars represent instances where only one model correctly classified the case. Y-axis records are sorted in ascending order, where cases near the top of the figure were classified correctly by fewer models and cases at the bottom of the figure were correctly classified by all models. The white space towards the top of the figure represents cases which were not correctly classified by any model. The rightmost column represents both the number of correct predictions associated with

records in a given band as well as the proportion of cases with the specified degree of consensus.

The lower rows of Figure 6 display the 68 cases on which complete and accurate consensus was reached by all models on attrition predictions. Considerable consensus was also reached on 23 cases, shown by the band where four models achieved accurate attrition predictions, however, the GNB optimized for F1 was unable to correctly classify most of these cases. The large white space within the GNB F1 column on the four- and three-vote bands shows substantial divergence from the decisions of the other models. The F1-optimized GNB also diverges, although to a lesser degree, in its correct predictions of cases in the one- and two-vote bands. As shown in Table 6, the F1-optimized GNB is a relatively conservative model with only 0.51 recall and relatively high precision (0.27). Better classification results can be obtained with other models, but the F1-optimized GNB's ability to identify some unique cases makes it a valuable in an ensemble.

The darkest bars in Figure 6 are associated with the F-Beta-optimized GNB. As detailed above, this model demonstrated the highest recall of any model on the hold-out set with 0.68, but the precision was also the lowest of any model at 0.18. The F-Beta-Optimized GNB clearly adds some unique predictive value to an ensemble, but this value must be taken into account alongside the model's error rate. Figure 7 displays the magnitude of false positives associated with all models. Here the false positive rate associated with the F-Beta-Optimized GNB is evident.

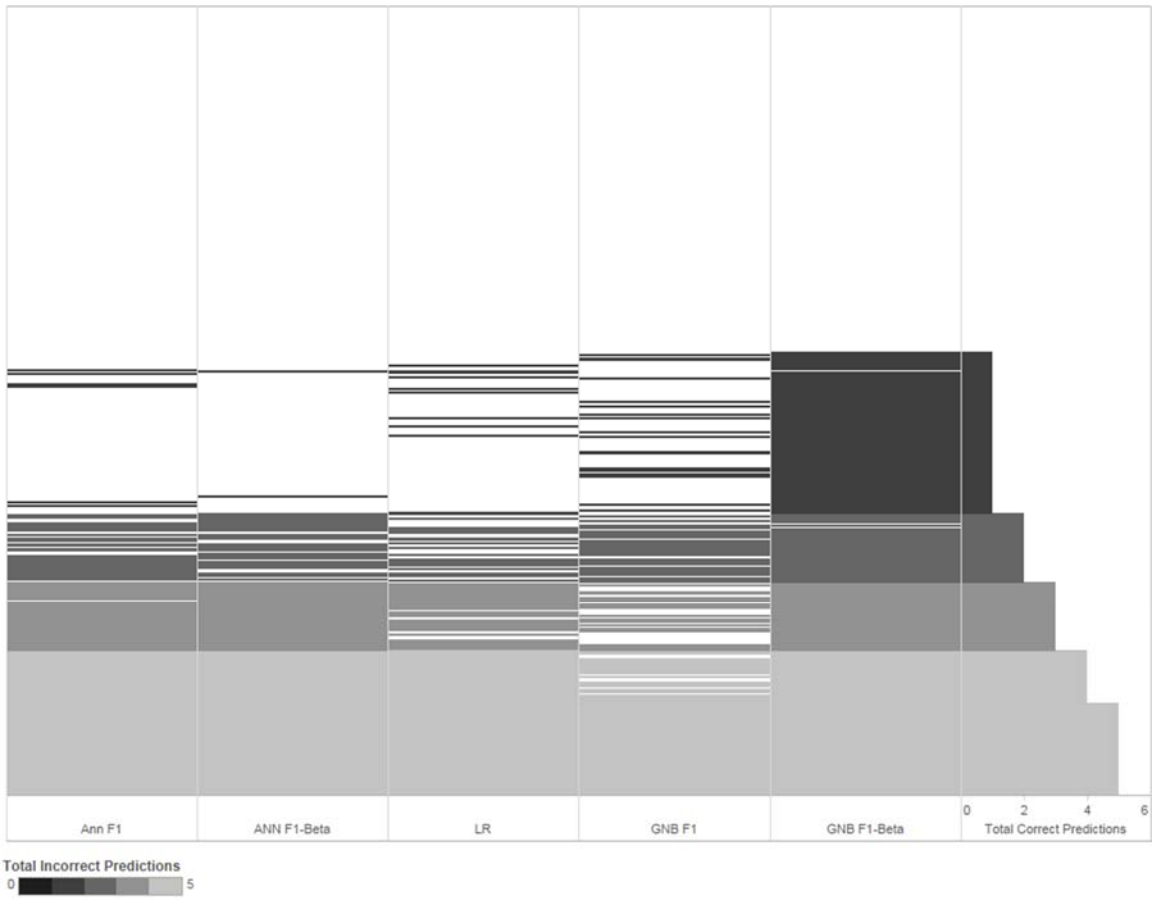


Figure 15. Incorrectly Classified Persistence Cases from Hold-Out by Model Type by Consensus

Figure 7 is similar to Figure 6 in terms of axes, although Figure 7 displays instances where the employed models made a false prediction of attrition. The white space towards the top of the figure represents correctly classified cases of persistence. Out of the 1,163 persistence cases all five models incorrectly classified 134 (11.52%) of these cases. Incorrect consensus was reached by four or more models on 211 (18.14%) cases. A majority of models made false classifications of attrition on 313 (26.91%) cases. There were 101 (8.68%) cases of persistence classified as attrition by two models, and as Figure 7 demonstrates, the GNB model optimized for F-Beta made a number of these. In fact, in

60 (59.41%) of the 101 cases where only two models made a false positive attrition prediction the GNB F-Beta-optimized model made an incorrect prediction. As the low precision and high recall of the F-Beta-optimized GNB suggests, a majority of the cases where only one model registered a false positive attrition prediction were attributable to the GNB optimized for F-Beta. Interestingly, the GNB optimized for F1 was able to correctly distinguish many of the cases where the other models incorrectly reached consensus on their attrition predictions. This can be seen as the white striations in the four and five total predictions bands in Figure 7. This highlights the ability of the F1-optimized GNB to correctly identify a greater proportion of persistence cases than any other individual model, as it exhibited a true negative rate of 79.02% on the hold-out dataset, compared to 76.53% for the logistic regression model and 71.45% for the F-Beta-optimized MLP. For a figure displaying case-level accuracy of persistence predictions by model see Appendix E.

Research question 3b. What is the profile of correctly classifiable versus incorrectly classifiable cases of attrition?

Correct classification of attrition was achieved by all models on 68 of the 177 cases, while none of the models were able to detect 36 cases of attrition, suggesting that some cases were more readily predictable than others. A basic descriptive analysis of item means for the more and less readily cases was conducted. Table 12 displays a selection of predictor means and standard deviations for hold-out set attrition cases that were predicted accurately by zero, one, four or five models.

Table 12

Feature Means of Accurately and Inaccurately Predicted Attrition Cases

Number of models with correct prediction	0	1	4	5
Number of cases	36 (inaccurate)	16	23	68 (accurate)
		M (SD)		
AGE_Q1EOT	18.28 (0.51)	18.31 (0.46)	18.48 (0.58)	18.28 (0.56)
First_Gen	0.22 (0.42)	0.19 (0.39)	0.09 (0.28)	0.13 (0.34)
GENDER	0.33 (0.47)	0.31 (0.46)	0.48 (0.5)	0.65 (0.48)
ADMINISTRATOR_RATING	3.81 (2.28)	4.56 (2.98)	5.57 (2.39)	6.87 (2.41)
SORGPAT_GPA	3.79 (0.26)	3.76 (0.27)	3.66 (0.29)	3.43 (0.38)
No_Aid	0 (0)	0.06 (0.24)	0.13 (0.34)	0.41 (0.49)
FM_APPLICATION_IND	0.61 (0.49)	0.75 (0.43)	0.74 (0.44)	0.62 (0.49)
ATHLETE_Q1EOT	0.14 (0.35)	0.06 (0.24)	0.04 (0.2)	0.01 (0.12)
Activity_Count_Q1	1.08 (0.95)	1.06 (0.83)	0.52 (0.77)	0.4 (0.71)
GREEK_Q1EOT	0.31 (0.46)	0.44 (0.5)	0.04 (0.2)	0.09 (0.28)
Intramurals_Activ_Type_Q1	0.47 (0.5)	0.5 (0.5)	0.35 (0.48)	0.24 (0.42)
LEP_Q1	0.03 (0.16)	0 (0)	0.09 (0.28)	0.18 (0.38)
Honors Program_Q1	0.11 (0.31)	0.06 (0.24)	0 (0)	0 (0)
Living & Learning Communities_Activ_Type_Q1	0.08 (0.28)	0 (0)	0 (0)	0 (0)
LLC: Leadership Program_Q1	0.08 (0.28)	0 (0)	0 (0)	0 (0)
Social_Activ_Type_Q1	0.31 (0.46)	0.44 (0.5)	0.04 (0.2)	0.09 (0.28)
HoldsNew_Q1	0.33 (0.67)	0.25 (0.56)	0.22 (0.51)	0.63 (0.78)
HOLD_COUNT_Q1	0.56 (0.68)	0.19 (0.39)	0.87 (0.68)	1.78 (1.64)
GRADE_VALUE_FSEM	87.78 (4.26)	85.56 (5.51)	86.22 (4.56)	73.72 (20.06)
W_Count_Q1	0.06 (0.23)	0.38 (0.6)	0.09 (0.28)	0.59 (0.94)
min_GradeVal_Q1	81.64 (5.31)	80.75 (7.5)	78.43 (4.27)	63.15 (16.48)
max_GradeVal_Q1	89.64 (1.16)	88.88 (2.6)	89.48 (1.06)	77.26 (17.64)
mean_GradeVal_Q1	86.85 (2.57)	85.35 (4.37)	84.45 (2.25)	70.71 (16.43)
Fail_Count_Q1	0 (0)	0 (0)	0 (0)	0.54 (1.08)
Proportion_belowCminus_Q1	0 (0)	0.02 (0.06)	0.01 (0.03)	0.38 (0.36)
CREDITS_EARNED_Q1	15.81 (1.41)	14.56 (2.5)	15.74 (1.33)	10.96 (5.04)
GPA_Q1	3.59 (0.3)	3.44 (0.45)	3.34 (0.24)	2.16 (1.13)
TRANSFER_HRS_Q1EOT	13.08 (14.18)	12.81 (17.99)	7.11 (11.03)	5.79 (14.58)
Academic_Difficulty_Q1	0 (0)	0 (0)	0.04 (0.2)	0.22 (0.64)
Cares_Submissions_Q1	0.08 (0.36)	0.06 (0.24)	0.3 (0.86)	0.57 (1.28)
Drugs_Alcohol_Q1	0.33 (1.03)	0 (0)	0.61 (1.71)	0.91 (1.95)
Mental_Health_Q1	0 (0)	0.06 (0.24)	0 (0)	0.21 (0.7)
ACT Comp Conv	27.92 (3.29)	27.56 (3.06)	27.43 (3.6)	26.54 (3.09)

ACT Math Conv	26.97 (3.81)	26.56 (2.65)	25.91 (4.86)	25.59 (3.36)
Total_Conduct_Q1	0.56 (1.5)	0.13 (0.48)	1.26 (2.47)	2.41 (4.09)
TOTAL_ACCEPT_AMOUNT	35,800.89 (17,368.65)	35,549.96 (21,130.79)	20,220.1 (18,268.83)	14,124.47 (16,630.8)
FIN_AID_TYPE_Scholarship	22,607.47 (10,187.98)	18,100.25 (10,641.05)	10,488.87 (7,983.98)	6,131.1 (7,492.7)
GIFT_OR_SELF_HELP_Gift Aid	29,042.83 (14,860.07)	26,106.44 (15,388.18)	12,860.7 (9,164.59)	9,967.94 (12,044.97)
FUND_SOURCE_DESC_Undergraduate Discount	25,373.78 (9,729.41)	19,134.94 (9,607.71)	11,724.83 (8,549.34)	9,062.81 (10,835.32)
FED_FUND_ID_PELL	822.78 (1,769.11)	726.88 (1,923.13)	644.57 (1,678.92)	444.15 (1,423.05)
IM_UNMET_NEED	-9,207.57 (17,327.2)	-12,186.89 (17,428.11)	694.9 (19,211.35)	2,188.44 (13,239.32)
RACE_DESC_White	0.83 (0.37)	0.88 (0.33)	0.91 (0.28)	0.93 (0.26)
ETHN_CDE_DESC_Not Hispanic or Latino	0.94 (0.23)	0.94 (0.24)	0.91 (0.28)	0.88 (0.32)
CITZ_CODE_Y	0.97 (0.16)	1 (0)	1 (0)	0.99 (0.12)
State_out-of-state	0.72 (0.45)	0.63 (0.48)	0.96 (0.2)	0.74 (0.44)
COLLEGE_1_Q1EOT_BUS	0.42 (0.49)	0.13 (0.33)	0 (0)	0.04 (0.21)

Note. Number of models with correct prediction = the total number of models where a correct attrition prediction was achieved; Number of cases = the number of cases on which the degree of consensus was reached; M(SD) = mean and standard deviation

As Table 12 demonstrates, a greater proportion of the students were first-generation college students for cases where zero or one model made an accurate attrition classification than for cases where four or more models made an accurate prediction (22% and 19% versus 9% and 13%, respectively). A smaller proportion were males in the zero or one prediction cases (33% and 31%, respectively versus 48% and 65% for the four and five prediction cases, respectively). The table also demonstrates that the less readily predicted cases had higher high school GPAs, higher minimum first quarter grades, higher average first quarter grades, lower rates of course failure, higher transfer credit hours, less incidence of academic difficulty, fewer behavioral care submissions, less drug or alcohol violations and higher composite and math ACT scores. When compared to the more readily predicted cases, a greater proportion of the unpredicted or singularly predicted cases were receiving aid, participating in athletics, belonged to

Greek organizations, participated in intramurals, belonged to the honors program, and attended the college of business. Total amount of accepted financial aid, scholarship aid, and gift aid were also higher for the less readily predicted cases. The cases from the unpredicted or singularly predicted group also accepted a slightly higher amount of federal Pell Grant aid, had a substantially higher average undergraduate discount and had much less unmet need than the readily predicted groups.

Overall, the less readily predicted cases tended to have scores, values and attributes that, when considered alongside important model predictors and associated coefficients, would suggest decreased odds of attrition. This finding was expected, as few or no models were able to classify these cases correctly. However, a greater proportion of students were first-generation college students and/or athletes in the less readily predicted groups. Also, a slightly smaller proportion of these students were white and a slightly larger proportion were from in state. The groups of readily and less readily predicted cases demonstrated some distinct differences in terms of central tendency on important features. The potential implications of these differences will be discussed further in the following chapter.

Research question four. Which neural network architecture results in the best classification performance?

Both the F-Beta- and F1-optimized MLP models exhibited better performance with the logistic activation function than with the hyperbolic tangent activation function. The F-Beta-optimized model utilized 75 predictors, one hidden layer with 50 neurons and an alpha regularization parameter of 0.01. The F1-optimized model utilized 75

predictors, two hidden layers with 25 neurons in each layer and a regularization parameter of 0.001.

The two models performed nearly identically on F-Beta (M = 0.47, SD = 0.03) and F1 (M = 0.34, SD = 0.02) metrics during the cross-validation phase, with each exhibiting very slight superiority on its respective metric. During cross-validation the F-Beta model demonstrated higher recall (M = 0.63, SD = 0.05) than the F1-optimized model (M = 0.61, SD = 0.04) while the F1 model exhibited higher precision (M = 0.24, SD = 0.02) than the F-Beta model (M = 0.23, SD = 0.02). ROC AUC was the same for both models in cross-validation (M = 0.71, SD = 0.03), while the overall accuracy was slightly better for the F1-optimized model (M= 0.69, SD = 0.02) than the F-Beta-optimized model (M= 0.67, SD = 0.03). During the final model evaluation on the hold-out set, both models achieved F-Beta scores of 0.48, but the F-Beta-optimized model achieved a higher F1 score of 0.36 than the F1-optimized model's score of 0.35. Overall accuracy of 0.70 and ROC AUC of 0.67 was achieved by both models during the final evaluation. Additionally, when looking at the final classification results, displayed above in Table 3, both MLPs made the exact number of correct and incorrect classification on the attrition group from the held-out dataset.

Notably, while the models made the same number of correct and incorrect predictions for the attrition group, they did not fully concur on a case by case basis. In the final evaluation, however, the F-Beta-optimized model was able to classify six additional cases of persistence correctly over the F1-optimized model, resulting in a true negative rate of 71.45% for the F-Beta model versus 70.94% for the F1-optimized model. The final evaluation metrics of these models were remarkably similar. The similarity in

terms of performance in both cross-validation as well as hold-out evaluation demonstrates the ability to achieve similar predictive functions and model performance with different MLP architectures.

The similarity of the two models can be seen in the final evaluation ROC curves of the two models. The similar trajectories of the models across decision boundaries demonstrates the average concurrence of the models, while it can be seen that either of the models could slightly outperform the other at varying decision points. Overall, however, the F-Beta model demonstrates slightly superior performance on the hold-out set.

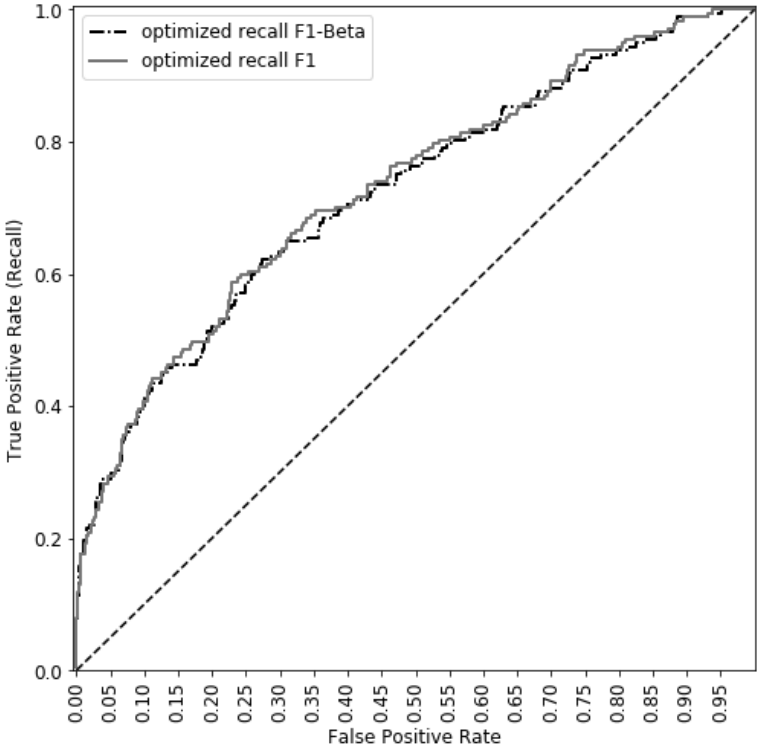


Figure 16. Final Evaluation ROC Curves: F1-Optimized and F-Beta-Optimized MLP Models

Finally, the generalizability of both MLP models should be noted. Each model demonstrated similar if not better metrics on the evaluation dataset than on the cross-validation dataset. This consistency in performance between training and final testing data is evidence that the models were not overfit and that they are able to perform well on unseen data.

CHAPTER FOUR: DISCUSSION

Summary of the Study

The current study leveraged conventional and less commonly examined student data in order to compare three different classification algorithms on their ability to model one-year persistence of college students using data from the first term of enrollment. Specifically, multilayer perceptron neural network and Gaussian naïve Bayes were compared against the most commonly used technique, logistic regression. The study utilized a grid search process in order to optimize dimensionality reduction and parameter tuning and to deal with class imbalance. Models were optimized using 10-fold cross-validation. Final models were trained on the entire cross-validation dataset and final evaluation was performed on a held-out dataset consisting of a single cohort. Each model was built twice, optimizing for F-Beta and F-1 scores. A majority voting ensemble and an empirically generated ensemble model were also created, and the results were compared against the best performing individual model. Models were primarily evaluated on F-Beta and F-1 results, but ROC-AUC, recall, precision, specificity and accuracy were also considered.

The results of the analysis revealed that a single logistic regression model performed optimally during cross-validation on F-Beta and F-1 scores. Both the MLP and GNB optimization processes resulted in different models when optimizing for F-Beta

and F-1. The logistic regression model outperformed both iterations of the alternative approaches, MLP and GNB, on F-Beta and F-1 for the final evaluation set. Both MLP models as well as the F-Beta-optimized GNB models outperformed the logistic regression model slightly on recall, although the trade-off in precision was too large and resulted in diminished F-Beta and F-1 scores.

The logistic regression model was compared to a majority voting ensemble in which predictions from all five models were given equal weight in a consensus-based prediction. The individual logistic regression model outperformed this model on F-Beta and F-1 scores as well as on precision, ROC-AUC, and accuracy. The logistic regression model was also compared to an empirically-derived weighted ensemble model in which the optimal model selection and weighting scheme was based on cross-validation results. The empirical ensemble was built twice, optimizing for F-1 and F-Beta. The F-Beta-optimized ensemble resulted in the F-Beta-optimized MLP and the logistic regression model being selected with weights of 2.00 and 3.00, respectively. The F-1-optimized ensemble selected the F-Beta-optimized MLP and the logistic regression, with weights of 1.00 and 3.00, respectively. Although cross-validation results suggested enhanced performance for both empirically-derived ensembles through the incorporation of the F-Beta-optimized MLP, the individual logistic regression model outperformed both empirical ensembles on F-Beta, F-1, precision, ROC-AUC, and accuracy when applied to the held-out evaluation set. Both the majority voting ensemble and the F-Beta-optimized empirical ensemble outperformed the individual logistic regression model slightly on recall, although, as with the individual models, the resulting trade-off in precision was too large and resulted in diminished F-Beta and F-1 scores for the ensembles.

Key Findings

Research question 1a. Which of the examined models results in the most desirable classification of students into attrition and persistence groups?

Research question 1a examined which of the proposed models resulted in the most desirable classification of students into attrition and persistence groups. Evaluation metrics as well as confusion matrices, ROC curves, and precision/recall curves were used to make the final determination of model performance, with F-1 and F-Beta scores serving as the primary means of evaluation and comparison. The logistic regression model demonstrated superior performance compared to both the F1- and F-Beta-optimized GNB and MLP models. In regard to final hold-out set evaluation the logistic regression model predicted 61.02% of attrition cases correctly with a precision rate of 23.47%. As noted in the previous chapter, as well as in the summary of the study section, a single logistic regression model performed optimally for F-Beta and F-1 scores during cross-validation. When applied to the held-out evaluation data, the logistic regression model attained higher F-Beta and F-1 scores than any other individual model. The model also resulted in the highest ROC AUC and precision scores observed from any model. ROC curves showed distinct separation between the logistic regression model's curve and the curves of the MLP and GNB models.

The results of this study support the use of logistic regression as a predictive modeling technique within the context of college student persistence (Gansemer-Topf, et al., 2014; Glynn, Sauer, & Miller, 2006; Lopez-Wagner, Campbell, & Mislevy, 2012; Miller (1999); Robb, Moody, & Abdel-Ghany, 2012). The results also emphasize the potential enhancements that can be made to such models by optimizing model parameters

for the scoring metrics of concern through a grid search process, as it allows all evaluated models to be precisely tuned during cross-validation in order to facilitate the best possible model performance on the held-out evaluation data. The use of cross-validation is also noteworthy, as it allows for the creation and selection of models with the highest potential generalizability to new data. The generalizability resulting from this technique can indeed be seen when comparing average cross-validation scores for each model to those models' scores during hold-out evaluation. As Tables 6 and 7 in Chapter Three demonstrate, model evaluation metrics were very similar from cross-validation to evaluation, with many of the models achieving higher metrics on the unseen hold-out data. In particular, logistic regression achieved higher F-Beta, F-1, precision, and accuracy scores on the hold-out set than it did on average in cross-validation.

Within the context of studies such as Dreiseitl and Ohno-Machado (2003), where the authors found better performance by artificial neural networks than by logistic regression models in 51% of reviewed studies, no difference between the two approaches in 42% of studies, and better performance by logistic regression in 7% of studies, the findings are not necessarily unexpected. While the literature suggests that neural networks tend to outperform logistic regression models on binary classification tasks, the efficacy of each approach is context-dependent. The results of the current study showed clearly superior performance by logistic regression, although both iterations of MLP neural network performed similarly to one another and had ROC curves closely approaching that of the logistic regression model.

The finding that logistic regression outperformed the alternative approaches on the persistence classification task diverges from findings in a methodologically similar

study by Delen (2011), in which the author found better results from an artificial neural network than from logistic regression. It should be noted, however, that Delen's results were based on cross-validation predictions from under-sampled student data. The final comparison and evaluation of classifier performance was not made based on model performance on a hold-out set. Additionally, Delen did not conduct parameter optimization based on *a priori* classification metrics. Furthermore, final model evaluation was based solely on the sensitivity and specificity rates. Notably, Delen's study under-sampled students who persisted in order to achieve a one to one class balance in the data, but the author failed to apply final trained models on an actual set of imbalanced data to assess the generalizability of the models. Delen's study found much higher rates of sensitivity (proportion of accurately predicted attrition cases) than of specificity (proportion of accurately predicted persistence cases), which calls into question the potential generalizability of the model, especially to the under-sampled and less distinguishable class of persisting students. The current study used synthetic and weight-based balancing for parameter optimization, but generalizability was established by using final trained models to make predictions on a set of unbalanced data. Additionally, the current study more strongly aligns with practice, as final models are built from five sequential years of data and then applied to a held out set from the following year.

Dreiseitl and Ohno-Machado (2003) also note the relative simplicity of building and interpreting logistic regression models compared to neural networks. This observation was reinforced through the model building process of the current study. MLPs have more parameters than logistic regression models and are also more

computationally expensive to train. This complexity becomes apparent when implementing a grid search process to optimize parameters within cross-validation, as MLP training times can quickly grow large, especially when compared to the time it takes to train a simpler model such as logistic regression or GNB.

It is important to note that the application of logistic regression in this study was characterized by attributes common to many practical machine learning modeling tasks. Statistical significance of included predictors was not a concern of the study, instead maximizing relevant scoring metrics through an empirical selection of variables was the goal. The use of an automated pipeline for data transformation, balancing, dimensionality reduction, and parameter tuning are all hallmarks of a machine learning approach. While the best performing model does offer some explanatory insights regarding persistence, the ultimate goal was to search the parameter space for each model in order to achieve the most efficacious and generalizable models built within the confines of a real-world modeling context.

Research question 1b. Does the ensemble model surpass the best performing individual model in terms of the examined classification metrics?

Research question 1b asked whether an ensemble model created from the individual models surpassed the best performing individual model on the examined scoring metrics. As detailed in the summary of results as well as in the results section, this question was addressed through the creation of three ensemble models. The first model was a majority rule voting (MVE) ensemble in which all final classifiers were given equal weight regarding the final prediction. Since model weighting was established *a priori*, there was no need for cross-validation in order to optimize the

weighting scheme. The second model was optimized for F-Beta and the third model was optimized for F-1. Cross-validation was used to develop the weighting scheme for the second and third models. Within the cross-validation process model weights ranging from zero to six were considered for the logistic regression model, the F-Beta-optimized MLP and both GNB models. All weighting permutations were assessed, and the weighting scheme with the best average results for the scoring metric of concern was adopted. No GNB models were empirically selected for inclusion in either the second or third ensemble. The logistic regression model was assigned a weight of three in both empirical ensembles, with the F-Beta-optimized MLP receiving a weight of two for the F-Beta-optimized ensemble and a weight of one for the F-1-optimized ensemble.

Both empirical ensembles performed slightly better on their respective optimization metrics than did logistic regression alone during cross-validation, hence the observed weighting schemes. However, during final evaluation on the hold-out set, logistic regression resulted in higher scores on F-Beta and F1 than any of the three ensembles. The logistic regression model also achieved higher precision, ROC AUC, and accuracy scores than any of the ensembles. The logistic regression model's superior performance on the hold-out set is evidence of its generalizability and the fact that it outperformed ensembles with inputs from additional models, serves as additional evidence that logistic regression, in this case, performed better than any individual model, as inputs from other models only degraded the scoring metrics of concern for the individual logistic regression classifier.

It should be noted that the MVE and the empirical ensemble optimized for F-Beta both achieved higher recall scores than the individual logistic regression model (0.63 and

0.62, respectively versus 0.61), however, the necessary tradeoff in precision to achieve these scores resulted in excessive diminishment of F-Beta and F-1 scores. Nevertheless, as the case-level comparison of classifier accuracy in Figures 6 and 7 demonstrate, there is enough differential prediction of individual cases between models to warrant further investigation of the efficacy of ensemble modeling within this context. At this point, the exploration of ensemble modeling within a student persistence context is relatively unexplored. As noted above, the empirical weight optimization process only considered integer values from zero to six. This decision was a result of computational expenditures necessary to find optimal weights within a cross-validation process. Future studies may find better performance from ensemble modeling by searching a less constrained weight space and/or by implementing a grid search and weight optimization process concurrently.

Research question 1c. Do any of the alternative algorithms perform better than the logistic regression model in regard to the metrics of concern?

This question sought to address whether or not any of the F-Beta-optimized or F-1-optimized MLP or GNB models performed better than logistic regression model on F-Beta or F-1 scores when applied to the held-out evaluation set. The analysis used to answer research question 1a also provided sufficient evidence to draw conclusions regarding research question 1c, as the logistic regression model performed superiorly on F-Beta and F-1 scores during both cross-validation and final evaluation. As noted above, both iterations of the MLP models outperformed logistic regression on recall during hold-out evaluation, each classifying two more attrition cases correctly than the logistic regression model (110 versus 108), however, recall was not the primary scoring metric of

concern. Instead, the study was primarily concerned with achieving a balance between high recall and precision, hence the use of F-1 and F-Beta scores. As noted above, precision and ROC AUC were also higher for the logistic regression model during final evaluation than for any other model. Logistic regression achieved the second highest overall hold-out evaluation accuracy with a score of 0.74, with the highest score of 0.75 attributed to the most conservative model evaluated, F1-optimized GNB. While the F1-optimized GNB was effective at distinguishing cases of persistence, its recall was insufficient and resulted in lower overall F-Beta and F-1 metrics.

This finding supports the choice of logistic regression in college student persistence modeling studies such as Lopez-Wagner, Carollo, and Shindlecdecker, further demonstrating the efficacy of using logistic regression as a predictive modeling technique in college persistence contexts. Notably, the current study diverges from efforts such as those by Lopez-Wagner et al., in that the primary focus of the current research was to build and compare models for early detection of student attrition risks instead of to implement logistic regression as a technique to explain student attrition risks across the entire first year of college. Results of the current study explore and support the role of logistic regression as a predictive tool and not necessarily as an explanatory tool, although logistic regression's application for explanatory purposes in the social sciences is strong.

The results of the current study also to some extent counter empirical comparisons of classifiers, such as Caruana and Niculescu-Mizil (2006), in which artificial neural networks were shown to outperform logistic regression on binary classification tasks, although it is notable that Caruana et al. (2006) was focused on classification applications

outside the domain of college student persistence. The study does, however, further support the finding by Caruana et al. (2006) that logistic regression tended to outperform naïve Bayes.

The finding that logistic regression outperformed both MLP models as well as all evaluated ensemble models suggests that the available data were best modeled by a linear model. The fact that both MLP models achieved similar results between each other and also similar, yet inferior, results to the logistic regression classifier, further suggests the appropriateness of a linear model within this context. The similarity between the MLPs and the logistic regression model was not surprising, as MLPs are universal function approximators. However, this result suggests that given the available data, the ability of the MLPs to produce non-linear solutions was not needed, as the best results were obtained using the linear logistic regression classifier. Given simpler linearly-solvable binary classification problems, it is likely that logistic regression can obtain better or at least similar results to MLPs. Given more complex or unstructured classification problems it is likely that the elasticity of MLPs, in terms of function approximation, would result in better performance of those classifiers when compared to logistic regression.

Research question 2a. Which features or composite features result in the best performing classification model?

This question sought to identify the most impactful features used in the best performing classification model. Since the logistic regression model performed best on the designated scoring metrics, the features used in that particular model were reviewed in order to answer this question. As noted in the previous chapter, a univariate feature

selection method, select K best, was used in order to identify the best potential predictors during cross-validation. The K features which resulted in the best cross-validation F-Beta and F-1 metrics were then selected for use in the final model. Notably, during cross-validation all models performed more optimally on F-1 and F-Beta when using select K best for dimensionality reduction versus PCA.

The 75 predictors used in the logistic regression model were ranked based on the absolute value of the product of their coefficient and standard deviation. This process is typically done to mitigate the influence of differential variable scaling on the ranking process, and although all numerical features were normalized, the categorical features were dummy coded. In order to control the importance of the dummy coded feature coefficients, the standard deviations of these variables were used to adjust the coefficient for ranking purposes. The top 40 predictors based on this methodology are displayed in Table 11, while Appendix G contains the full list of 75 ranked features.

As discussed in the preceding chapter, four of the top ten predictors were related to financial aid, four were related to academic performance, one was related to pre-collegiate qualifications, and one was related to student conduct. Features with ranked importance between 11 and 20 consisted of financial aid figures, student account holds, conduct violations, academic performance, and student activity counts. Several of the categorical features began to appear in features with ranked importance between 21 and 40, with specific residence halls, sophomore class standing, out-of-state student status, and primary affiliation with the college of business emerging as important predictors within this band. Overall, numerical features emerged as better predictors through the

employed ranking methodology, with categorical features clustering towards the lower end of the ranked predictors.

As discussed in the results, the top 20 predictors consisted of a mix of traditional predictors, such as overall financial aid figures and academic performance data, as well as less commonly leveraged data such as counts of student conduct violations and computed metrics of academic performance not commonly used by the institution. Award types disaggregated by fund source also emerged as useful predictors. The identification of predictive value in both the standard financial aid fields and the institution-specific disaggregates support the findings of research regarding the role of financial aid in persistence, such as St. John, Cabrera, Nora, and Asker (2000).

The results of the current research, to some extent, also support the predictive value of features such as GPA and ACT scores which have been found to be important predictors in many persistence studies, such as Stewart, Lim and Kim (2015). However, in the current study the coefficients from the logistic regression model associated with GPA, ACT composite, and ACT math scores were all negative, with odds of 1.11, 1.09, and 1.09, respectively, these results suggest that increased scores on these features lead to greater odds of attrition. This finding is interesting in that it is contrary to typical trends in persistence research regarding the impact of GPA and test scores. Notably, however, the unweighted mean grade value predictor, which was very similar to the first quarter GPA feature, was the fifth best overall predictive feature in the best model with associated odds of 0.81. The first quarter GPA and mean GPA variables also have the same relatively large negative correlations within the cross-validation set at $r = -0.23$.

Further study should be conducted in order to identify the impact of including two such similar predictors in a persistence model.

Another interesting observation based on variable ranking and coefficients was that the count of total conduct violations emerged as one of the top ten predictors, with a coefficient suggesting that greater numbers of conduct violations are associated with greater odds of persistence. The remaining conduct violations, with the exception of drug and alcohol violations, suggested the opposite trend, in which greater counts of conduct in those areas were associated with increased odds of attrition. Notably, most conduct cases overall are attributable to drug and alcohol violations. These observations concur with the findings of Martinez, Sher, and Wood (2008), in which alcohol use was found to be positively associated with persistence. This observation suggests the potential role of social engagement within drinking related activities on college campuses. The relationship between alcohol consumption and attrition should be investigated more thoroughly in future studies.

Overall, the most important predictors from the best fitting model, logistic regression, aligned with the broad categories of predictors identified by Ishler and Upcraft (2004). Notably, additional predictive value was found in disaggregates or derivatives of commonly used predictors. Additionally, less commonly employed predictors such as place of residence, student conduct violations and student accounts information also emerged as important predictors of attrition.

Research question 2b. Which of the rarely explored data elements are the most powerful predictors?

This question sought to determine which of the rarely explored data elements emerged as the most important predictors in the best fitting model. The methodology for addressing this question was the same feature ranking process as described for research question 2a. As discussed in the preceding chapter, the most important of the rarely explored predictors from the best fitting model included the disaggregated financial aid figures (FUND_SOURCE_DESC_Undergraduate Discount, FUND_SOURCE_DESC_State Funding, FIN_AID_TYPE_Scholarship, FUND_SOURCE_DESC_Departmental Funded Schl), the student account hold features (HoldsNew_Q1, Active_Holds_Q1), the custom computed figures based on academic performance data (Proportion below Cminus Q1, mean_GradeVal_Q1, min_GradeVal_Q1), the housing locations variables (BUILDING_DESC_Q1_McFarlane Hall, BUILDING_DESC_Q1_Centennial Towers), the conduct features (Total_Conduct_Q1, Cares_Submissions_Q1, Sexual EEO Q1, Endangerment_Weapons_Provoke_Q1, Academic_Difficulty_Q1, Dishonesty_Q1, Drugs_Alcohol_Q1, Mental_Heallth_Q1), and the involvement indicators (Activity Count Q1, Living & Learning Communities_Activ_Type_Q1, social_Activ_Type_Q1, Honors Program Q1).

Overall, these predictors were intended to serve as proxies for social and academic engagement or to allow for further nuance in terms of commonly used predictors. For example, disaggregating the total institutional award by fund sources, allowed for a more specific analysis of attrition as it relates to award magnitudes within specific categories. As colleges collect and store more and different student data attributes, these information sources should be used in the attrition modeling process in

order to continually identify valuable sources of predictive information. The fact that the best performing model, logistic regression, and the second and third best performing models, F-Beta-optimized MLP and F-1-optimized MLP, all performed similarly on recall, 0.61, 0.62, and 0.62, respectively, suggests that perhaps a ceiling has been reached in terms of early detection of attrition given the available information. Identifying the antecedents of attrition for the less discernable cases discussed in regard to research question 3b could potentially result in additional sources of novel information to aid in the modeling process.

Research question 3a. To what extent do the models differ in terms of the cases that they accurately classify or misclassify?

This question sought to identify the level of consensus between the five evaluated individual models on predictions for the 177 cases of attrition in the hold-out evaluation set. In order to address this question, case-level consensus was calculated across all five models. The results indicated that all five models reached accurate consensus regarding attrition on 68 (38.43%) of the attrition cases. Four or more models reached consensus on 91 (51.41%) cases, and three or more models made accurate classifications of attrition on 113 (63.84%) cases. Only 16 (9.04%) of the cases were accurately classified by one model. None of the models were able to accurately classify 36 (20.34%) cases.

In order to further investigate the degree of consensus between models a visualization technique, case-level classifier consensus density plot (C3-DP), was devised. These plots can be viewed in Figures 6 and 7, with an additional iteration in Appendix E. The C3-DPs express the level of consensus between models on specific cases, with lighter shades representing more readily classifiable cases and darker shades

representing less readily classifiable cases. The C3-DPs are used in the preceding chapter to depict correct and incorrect consensus. Since the C3-DPs represent consensus at a case-level, white space in the figures can be interpreted as instances where a specific classifier did or did not make a relevant prediction. At this time, such an approach to comparing case-level classifier performance does not appear to be present in any of the literature regarding empirical comparisons of classifier performance.

The C3-DPs presented in the results demonstrate the high level of consensus on the 68 more readily discernable cases and show the inability of the most conservative model, the F1-optimized GNB, to distinguish cases on which most of the other models reached correct consensus. The model also shows the unique predictive value of the least conservative model, F-Beta-optimized GNB. Relatively novel correct attrition predictions made by the logistic regression model are also distinguished. The C3-DPs further demonstrate the similarity between the F1-optimized and F-Beta optimized MLP models, highlighting the fact that while the models made the same number of correct attrition predictions, they did not reach full consensus at a case-level. This visualization technique has utility in future classifier comparison studies and could potentially aid in the creation of ensemble models, as it highlights case-level strengths and deficiencies of individual classifiers.

Research question 3b. What is the profile of correctly classifiable versus incorrectly classifiable cases of attrition?

This question addressed the profile, in terms of utilized data, of correctly versus incorrectly classifiable cases of attrition. The question was intended to provide further insight into the cases where evaluated models were unable to make accurate attrition

predictions. As mentioned in regard to research question 3a, none of the evaluated models made accurate predictions on 36 (20.34%) of the attrition cases, while only one model made an accurate prediction on 16 (9.04%) of the cases. Attrition cases were successfully classified by all models in 68 (38.42%) cases and correctly classified by four models in 23 (12.99%) cases. In order to examine the general differences between scores, values, and attributes of the less readily predicted cases (those predicted by zero or one model) and the more readily predicted cases (those predicted accurately by four or more models), means and standard deviations for relevant variables were calculated. These figures are displayed in Table 12 in the preceding section.

The results of this descriptive analysis suggest that, in general, models had a more difficult time discerning cases with higher pre-collegiate academic qualifications, greater first quarter academic performance, fewer account holds, fewer conduct violations, fewer alcohol-related conduct violations, lower unmet need, greater numbers of scholarship funds, greater involvement in activities, and proportionally greater athletics participation. Additionally, a proportionally greater amount of the less readily classifiable cases were non-white, female, in-state applicants, and first-generation college students.

The classification difficulty for these cases makes sense when taking into account the coefficients from the logistic regression model, since that model tended to ascribe decreased odds of attrition to students with higher scores (such as those exhibited by these groups) on many of the metrics reviewed in the descriptive analysis. This analysis highlights multiple potential directions for future research. First, it is possible that there may be something unique about these groups that is otherwise not being captured by the data used in the model. For example, there could be some specific social processes

which disproportionately impact female, in-state, first-generation athletes who engage in fewer alcohol-related social scenarios. A study concerning students with these general traits over the course of the entire first year of college could potentially illuminate the attrition process within this group. Second, a qualitative follow-up regarding this group could potentially be conducted to learn more about the factors influencing the attrition process. Finally, the fact that such a large proportion of attrition cases were not accurately predicted by any model, suggests the potential need to expand the modeling problem from a binary classification problem to a multinomial classification task. In order to facilitate such a modeling effort, it would first be necessary to expand the possible classes of outcomes. This could potentially be done by incorporating a third class to represent students who transfer out of the institution. In order to accomplish this, data sources and temporal constraints would need to be evaluated.

Research question four. Which neural network architecture results in the best classification performance?

This question addressed, independent of the performance of the other evaluated classifiers, which MLP architecture resulted in the best overall classification performance. In order to address this question classifier performance was evaluated based on cross-validation performance as well as on final hold-out set evaluation. The primary scoring metrics of concern, F-Beta and F-1, were utilized in this analysis as well as all other available scoring metrics.

As mentioned in preceding chapters, a grid search process was utilized when searching the parameter space for model specifications that optimized the scoring metric of concern. Two final MLP models were built, one optimizing for F-Beta and the

other optimizing for F-1. The F-Beta-optimized model utilized 75 predictors, an alpha regularization parameter of 0.01, the logistic activation function, and a single hidden layer with 50 neurons. The F-1-optimized MLP utilized 75 predictors, an alpha regularization parameter of 0.001, the logistic activation function, and two-hidden layers with 25 neurons each. Within the grid search process, the logistic activation function was assessed against the hyperbolic tangent activation function, as described in the methods. The fact that both models achieved optimal scores utilizing the logistic function is counter to the observation by Karlik and Olgac (2010) in which the researchers found slightly higher classification accuracy with a hyperbolic tangent function than with logistic sigmoidal activation functions.

Both the F-Beta- and F-1-optimized MLPs performed slightly better than one another on their optimization metrics of concern in cross-validation, although the F-Beat-optimized MLP performed better on F-1 score during final hold-out evaluation than did the F-1-optimized model. The models scored nearly identically on every other scoring metric of concern during final evaluation. A visual inspection of ROC curves suggested very similar trajectories between the two models in terms of the tradeoff between recall and FPR across all decision thresholds. Based on evaluation metrics F-Beta and F-1, however, the F-Beta-optimized architecture outperformed the F-1-optimized architecture. Additionally, the value of the F-Beta model over the F-1 model was further evidenced by the fact that this model was empirically selected for inclusion in the second and third ensembles, whereas the F-1-optimized model was not selected for inclusion. Notably, both MLP models achieved greater hold-out recall than did the logistic regression model, classifying 110 of 177 cases of attrition accurately versus the 108 cases classified

correctly by the logistic regression model. However, the tradeoff in precision was too large and diminished overall F-Beta and F-1 metrics. Overall, the results regarding MLP architecture support the notion that very similar predictive performance can be achieved between MLP models utilizing different architectures.

Contributions and Implications for the Field of Research Methods and Statistics

The results and analyses conducted in the current study make several contributions to the field of research methods and statistics. First, the process used for comparing classifiers demonstrates the importance of utilizing a holistic approach to evaluating classifier performance. Specifically, the current research outlined a process in which primary model evaluation metrics, F-Beta and F-1, were decided upon *a priori*. Then, models optimized for those metrics were compared in terms of their performance on a hold-out dataset. While models were primarily compared on F-Beta and F-1 performance, other metrics, such as recall, precision, ROC AUC, true negative rate, and accuracy were taken into account. Additionally, ROC curves and precision/recall plots were evaluated to assess model characteristics and to compare performance. Model generalizability from cross-validation to final evaluation was also considered. Finally, case-level model performance on the evaluation set was considered in order to assess differential predictive value and to gain a sense of model consensus.

The use of grid search to identify optimal parameters during cross-validation for each model is also important, as it encourages the best possible classification results for each model on the given evaluation metrics. Essentially, this process attempts to ensure that the best possible classifier will be built for each model type, allowing for fair

comparisons between classifiers. Any study comparing classifiers should first take steps to optimize those classifiers for the relevant evaluation metrics.

Additionally, the temporal ordering of balancing, normalization, and variable selection utilized in this study should be observed for future studies within the domain. The results of classification studies can be greatly biased when balancing and variable rescaling are done outside of individual cross-validation splits, as they introduce data leakage between training and testing folds.

The case-level classifier consensus density plot (C3-DP) that was built in order to address classifier consensus and disagreement was also a unique contribution to the field. Such an approach to comparing classifier performance at a case-level was not observed in the literature at the time of this study. This plot is an effective approach to assessing differential predictive value of classifiers and has potential utility in ensemble model creation, as it shows the overlap and divergence of classifier predictions.

The current research also highlights the potential utility of ensemble modeling and demonstrates two different approaches to identifying ensemble weighting schemes. While the best fitting individual model outperformed all ensemble models during final evaluation, it is worth noting that the empirically derived ensemble models included the use of the F-Beta-optimized MLP. When the goal of a study is to build the best classifier, an attempt should always be made to incorporate input from multiple models. An observation that an individual model outperforms and ensemble of models can also be used as further evidence of an individual classifier's superior performance compared to the other individual classifiers in the ensemble. In this regard, the current study

demonstrates that ensemble models can serve as both classification enhancement techniques and sources of evidence for individual classifier comparisons.

Finally, the relative performance of the models addressed in this study contribute to the current body of literature comparing classification techniques, as the results provide evidence of potential superior performance of logistic regression compared to classifiers such as MLP and naïve Bayes. This result is particularly relevant within a college student persistence context, and is also relevant to the greater body of literature regarding classifier comparisons.

Contributions and Implications for the Field of Institutional Research

The current study also has several implications for the fields of persistence and institutional research. First, the study demonstrates how Tinto's (1986) notion of an early alert system might be put into practice currently. The study outlines general categories and sources of relevant student data and demonstrates the ability to make useful predictions about one-year persistence after the first ten weeks of college. Furthermore, the analysis of important predictive features confirms many of the findings of previous college persistence literature, as these features align with the academic and social engagement predictors that are often referenced. Less frequently explored predictors, such as conduct, residential building assignment, and disaggregated financial aid information, were also shown to have predictive value in the modeling process.

The fact that the best fitting model from the current research attained recall of 61.02%, combined with the case-level consensus analysis, indicates that there are a number of cases that are difficult to classify. One potential avenue for addressing this is the expansion of the classification from binary to multinomial. Such an expansion could

allow for the modeling of additional outcomes that are not well-captured in the binary outcome categories of persist and attrit. Furthermore, the potential for multinomial classification also begs the question of whether the terms persist and attrit accurately represent the outcomes being studied. One could argue that some of the students in the attrition group are made to leave the university and are not necessarily leaving under their own will. Furthermore, it is likely that many of those who attrit within the first year transfer to different colleges due to a variety of reasons. These individuals do persist in the process of college education but do not persist at the institution of concern. Expansion of the outcome categories for this problem could result not only in better classification results but also in more accurate terminology related to the outcomes.

It is also worth noting that the majority of students from the research site used in the current study leave between the end of the last term of the first academic year and the beginning of the first term of the second academic year. In other words, most of the attrition takes place over the course of the first summer. This institutional attribute is important to note, as it has potential implications for model building and model implementation. This pattern of leaving is not necessarily common to all college settings, and it highlights the need to build and implement models within specific contextual circumstances.

Finally, it is important to note that while all models within the current study generalized well from cross-validation to final evaluation, there is no evidence of generalization from one institutional context to another. Additionally, given the nuances and variability in terms of the data collected by individual institutions, it is unlikely that the evaluated models from the current study would be able to be evaluated within another

context. Furthermore, it is unclear from the current study as to how the evaluated models would perform relative to one another if they were trained on smaller populations or even on modest samples. These observations further underscore the need for thorough context-specific modeling and the assessment of unique institutional characteristics that may inform modeling decisions.

Limitations

The current study adopted an *a priori* set of classification metrics for both optimization and final model evaluation and comparison. While this decision increased the veracity of the evaluation and comparisons made within the study it is still important to note that there is a degree of subjectivity involved in comparing binary classification approaches. While parametric and non-parametric tests exist for comparing classifiers (Desmar, 2006; Dietterich, 1998; Goodman, 1963), there is little agreement on the validity and appropriateness of these methods. Furthermore, formal tests for differences between classifiers are only conducted in a minority of classification comparisons (Caruana, 2006). The goal of the current research was not to identify statistically significant differences between classification approaches but was instead to use unique data in an applied context while leveraging parameter tuning processes in order to compare optimized classifier performance on specific, context-appropriate metrics in a real-world setting. There is abundant use of classification in industry and applied educational research, and in many respects the choice of a classifier within these contexts involves a multi-tiered inspection of scoring metrics, generalizability estimates and visual

representations of performance. Nevertheless, subjectivity can be identified to some extent in any approach to classifier comparison.

An additional limitation of the study can be found in the grid search process. When searching the parameter space for optimal model settings, the complexity and time to train grows exponentially, especially for MLPs which are more computationally expensive to train than the other evaluated models. With additional computing resources a larger parameter space could have been searched in order to find a truly optimal set of parameters for each model.

Recommendations for Future Research

This current study was designed to allow for a high level of reproducibility. It is important that future studies of classification, both within and outside the domain of college student persistence, take measures to ensure the reproducibility of their results. Great care was also taken in the current research to eliminate any potential data leakage. This threat becomes increasingly prominent in circumstances where balancing, normalization and transformation are not executed appropriately within cross-validation. When these steps are not performed correctly, such information leakage can greatly bias and inflate results.

Balancing within the current study was achieved through the use of SMOTE as well as through a class weight parameter for logistic regression. Future studies should examine the impact of different balancing techniques on both within-model performance and generalizability. The application of SMOTE, specifically, has been shown to result in more generalizable models by reducing overfitting (Chawla, Bowyer, Hall, &

Kegelmeyer, 2002). Since models in the current study generalized well to the hold-out set, an empirical investigation regarding the degree to which this attribute is enhanced through balancing would be useful.

The current research could also be extended in several logical ways. First, the ability to detect a relatively large proportion of students who will not persist for one year suggests the potential for building and applying models during the admissions process in order to act as screening mechanisms for students who are unlikely to stay at an institution. There is potentially an interesting area for expansion, however, it must be noted that basing admissions decisions on likelihood to persist is wrought with many potential ethical quandaries. For example, it is realistic that a model built using only pre-collegiate data generated during the admissions process might attribute increased likelihood of attrition to first-generation college students with greater financial need and fewer family resources. Giving students witch such characteristics less consideration during the admissions process would not only degrade the diversity of the incoming class in terms of background and experiences, but it would also be explicitly unethical. Research regarding the integration of persistence outcomes into the admissions process must be conducted carefully and framed in an intentional and fair manner.

A second logical extension of the research would be to model graduation outcomes as well as longer-term persistence intervals, such as two-year and three-year persistence. Such research would allow for the exploration and expansion of the set of predictors used in the current study. The longer students persist, the more data they generate, and it is likely that some of the features that demonstrated predictive value in the current study would be less valuable when modeling different outcomes and intervals

of persistence. Future studies may also want to address rank ordering students by model probability scores instead of simply labeling them by their predicted class. Such a rank ordering could be useful in implementations of an early alert system in which students predicted to leave the institution are banded in terms of risk level.

Finally, within the context of college student persistence, the predictive value of additional sources of novel information should continue to be assessed. Future classification efforts in this field would be well-served by identifying and leveraging data sources that better address student attitudes and experiences.

References

Abbott, D. *Applied Predictive Analytics: Principles and Techniques for the Professional Data Analyst*. Indianapolis, Indiana: John Wiley & Sons, Inc.

Association for Institutional Research. (2016). *Defining institutional research: Findings from a a national study of ir work tasks.* Tallahassee, FL: Lillibridge, F., Seing, R. L., Jones, D., Ross, L. E.

Astin, A. W. (1984). Student involvement theory: A developmental theory for higher education. *Journal of College Student Personnel, 25*(4), 297-308.

Borkar, S., & Rajeswari, K. (2014). Attributes selection for predicting students' academic performance using education data mining and artificial neural network. *International Journal of Computer Applications, 86*(10), 1-5.

Braunstein, A., McGrath, M., & Pescatrice, D. (2000). Measuring the impact of financial factors on college persistence. *Journal of College Student Retention, 2*(3), 191-203.

Carey, K. (2004). *A matter of degrees: Improving graduation rates in four-year colleges and universities* (A Report by the Education Trust). Retrieved from: http:// edtrust.org/wp-content/uploads/2013/10/highered.pdf

Cabrera, A. F., Nora, A., & Castaneda, M. B. (1993). College persistence: Structural equations modeling test of an integrated model of student retention. *The Journal of Higher Education, 64*(2), 123-139.

Campbell, C. M., & Mislevy, J. L. (2012). Student perceptions matter: Early signs of undergraduate student retention/attrition. *Journal of College Student Retention, 14*(4), 467-493. doi: http://dx.doi.org/10.2190/CS.14.4.c

Cao, L. J., Chua, K.S., Chong, W.K., Lee, H.P., & Gu, Q.M. (2003). A comparison of PCA, KPCA and ICA for dimensionality reduction in support vector machine. *Neruocomputing, 55*, 321-336. doi:10.1016/S0925-2312(03)00433-8

Caruana, R., & Niculescu-Mizil, A. (2006, June). *An empirical comparison of supervised learning algorithms.* Paper presented at International Conference on Machine Learning, Pittsburgh, PA. doi: 10.1145/1143844.1143865

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique, *Journal of Artificial Intelligence Research, 16*, 321-357. doi: 10.1613/jair.953

Delen, D. (2012). Predicting student attrition with data mining methods. *Journal of College Student Retention, 13*(1), 17-35. doi: 10.2190/CS.13.1.b

Desmar, J. (2006). Statistical comparisons of classifiers over multiple data sets. *Journal of Machine Learning Research, 7*. 1-30.

Dietterich, T. G. (1998). Approximate statistical tests for comparing supervised classification learning algorithms. *Neural Computation, 10*. 1895-1923.

Downey, A. (2013). *Think Bayes*. Sebastopol, CA, USA: O'Reilly.

Dreiseitl, S., Ohno-Machado, L. (2002). Logistic regression and artificial neural network classification models: A methodology review. *Journal of Biomedical Informatics, 35*. 352-359. doi: 10.1016/S1532-0464(03)00034-0

Du, K., & Swamy, M. N. S. (2014). *Neural Networks and Statistical Learning.* London: Springer.

Fan, L., & Poh, K. L. (209). Improving the naïve Bayes classifier. In J. R. R. Dopico, J.

D. de la Calle, & A. P. Sierra (Eds.), *Encyclopedia of artificial intelligence*. Hershey, PA: IGI Global. Retrieved from: http://search.credoreference.com/content/entry/igiai/improving_the_na%C3%AFve_bayes_classifier/0

Gansemer-Topf, A. M., Zhang, Y., Beatty, C. C., Paja, S. (2014). Examining factors influencing attrition at a small, private, selective liberal arts college. *Journal of Student Affiars Research and Practice, 51*(3), 270-285. doi: http://dx.doi.org/10.1515/jsarp-2014-0028

Glynn, J. G., Sauer, P. L., & Miller, T. E. (2006). Configural invariance of a model of student attrition. *Journal of College Student Retention, 7*(3-4), 263-281.

Goodman, L. A. On methods for comparing contingency tables. *Journal of the Royal Statistical Society, 126*(1). 94-108. Retrieved from: http:// www.jstor.org/stable/2982447

Harvey, A., & Luckman, M. (2014). Beyond demographics: Predicting student attrition within the bachelor of arts degree. *The International Journal of the First Year in Higher Education, 5*(1), 19-29.

Hepner, G. F. (1990). Artificial neural network classification using a minimal training set: Comparison to conventional supervised classification. *Photogrammetric Engineering and Remote Sensing, 56*(4), pp. 469.473.

IPEDS 2015-16 Data Collection System, Retention rate definition. Retrieved from: https://surveys.nces.ed.gov/ipeds/VisGlossaryAll.aspx

IPEDS 2015-16 Data Collection System, First-time undergraduate definition. Retrieved from: https://surveys.nces.ed.gov/ipeds/VisGlossaryAll.aspx

IPEDS 2015-16 Data Collection System, First-year undergraduate definition. Retrieved from: https://surveys.nces.ed.gov/ipeds/VisGlossaryAll.aspx

Ishitani, T. T. (2006). Studying attrition and degree completion behavior among first-generation college students in the United States. *The Journal of Higher Education, 77*(5), 861-885.

Ishler, J. L., & Upcraft, M. L. (2004). The keys to first-year student persistence. In M. Upcraft, J. Gardner, & B. Barefoot (Eds.), *Challenging and supporting the first-year student* (pp. 27-46). Indianapolis, IN: Jossey-Bass.

Jing, Y., Pavlovic, V., Rehg, J. M. (2008). Boosted Bayesian network classifiers. *Machine Learning, 73*, 155-184. doi: 10.1007/s10994-008-5065-7

Kambhatla, N., & Leen, T.K. (1997). Dimension reduction by local principal component analysis. *Neural Computation, 9*, 1493-1516. doi: 10.1162/neco.1997.9.7.1493

Karlik, B., & Vehbi, A. (2010). Performance analysis of various activation functions in generalized MLP architectures of neural networks. *International Journal of Artificial Intelligence and Expert Systems, 1*(4), 111-122. Retrieved from https://www.researchgate.net/profile/Bekir_Karlik/publication/228813985_Performance_Analysis_of_Various_Activation_Functions_in_Generalized_MLP_Architectures_of_Neural_Networks/links/004635229e9d608b3a000000.pdf

Kohavi, R. (1995). *A study of cross-validation and bootstrap for accuracy estimation and model selection.* Paper presented at the International Joint Conference on Artificial Intelligence, Montreal, Quebec Canada.

Kuh, G. D., Kinzie, J., Buckley, J. A., Bridges, B. K., & Hayek, J. C. (2007). *Piecing*

together the student success puzzle: Research, propositions, and recommendations (ASHE Higher Education Report, 32(5).) doi: 10.1002/ aehe.3205

LeCun, Y. Who is afraid of non-convex loss functions? [Presentation]. Retrieved from: https:// www.cs.nyu.edu/~yann/talks/lecun-20071207-nonconvex.pdf

Lopez-Wagner, M. C., Carollo, T., & Shindledecker, E. *Predictors of Retention: Identification of Students At-Risk and Implementation of Continued Intervention Strategies*. Retrieved from: http://ir.csusb.edu/documents/PredictorsofRetention_ 002.pdf

Martinez, J. A., Sher, K. J., & Wood, P. K. (2008). Is heavy drinking really associated with attrition from college? The alcohol-attrition paradox. *Psychology of Addictive Behaviors, 22*(3), 450-456. doi: 10.1037/0893-164X.22.3.450

Marsland, S. (2014). *Machine learning: An algorithmic perspective*. New Jersey, USA: CRC Press. Retrieved from: https://ebookcentral.proquest.com/lib/du/detail. action?docID= 1591570

McCormick, A. C. (2003). Swirling and double-dipping: New patterns of student attendance and their implications for higher education. *New Directions for Higher Education, 121*, 13-24.

Menard, S. (2010). *Logistic regression: From introductory to advanced concepts and applications.* Thousand Oaks, CA: SAGE Publications. doi: 10.4135/ 9781483348964

Milem, J. F., & Berger, J. B. (1997). A modified model of college student persistence:

Exploring the relationship between Astin's theory of involvement and Tinto's theory of student departure. *Journal of College Student Development, 38*(4), 387-400.

Miller, T. E., Tyress, T. M. (2009). Using a model that predicts individual student attrition to intervene with those who are most at risk. *Educational and Psychological Studies Faculty Publications, 28*, 13-19.

Pascarella, E. T., & Terenzini, P. T. (1979). Student-faculty informal contact and college persistence: A further investigation. *The Journal of Educational Research, 72*(4), 214-218. Retrieved from: http://www.jstor.org/stable/27537224

Pereira, F., Mitchell, T., & Botvinick, M. (2009). Machine learning classifiers and fMRI: A tutorial overview. *NeuroImage, 45*, 199-209. doi: 10.1016/j.neuroimage. 2008.11.11007

Riedmiller, M. (1994). Advanced supervised learning in multi-layer perceptrons: From backpropagation to adaptive learning algorithms. *Computer Standards & Interfaces, 16*(3), 265-278. doi: 10.1016/0920-5489(94)90017-5

Rish, I. (2001). *An empirical study of the naïve Bayes classifier*. T.J. Watson Research Center. Retrieved from: https://www.researchgate.net/publication/228845263_ An_Empirical_ Study_of_the_ Naive_Bayes_Classifier

Robb, C. A., Moody, B., & Abdel-Ghany, M. (2012). College student persistence to degree: The burden of debt. *Journal of College Student Retention, 13*(4), 431-456. doi: http:// dx.doi.org/10.2190/CS.13.4.b

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature, 323*(9), p. 533-536.

Sainani, K. L. (2014a). Explanatory versus predictive modeling. *Physical Medicine and Rehabilitation, 6*(9), p. 841-844. Retrieved from http://dx.doi.org/10.1016/j.pmrj.2014.08.941

Sainani, K. L. (2014b). Logistic Regression. *Physical Medicine and Rehabilitation, 6*(12), p. 1157-1162. Retrieved from http://dx.doi.org/10.1016/j.pmrj.2014.10.006

Scikit-Learn (Neural Networks Supervised) [Computer software]. Retrieved from: http://scikit-learn.org/stable/modules/neural_networks_supervised.html

Shanmuganathan, S. (2016). Artificial neural network modelling: An introduction. In S. Shanmuganathan, & S. Samarasinghe (Eds.), *Artificial Neural Network Modelling* (pp. 1-14). Switzerland: Springer.

Shmueli, G., Patel, N. R., & Bruce, P. C. (2016). *Data mining for business analytics.* Hoboken, New Jersey, USA: John Wiley & Sons. Retrieved from: https://ebookcentral.proquest. com/lib/ca/detail.action?docID=4526672

John, E. P., Cabrera, A. F., Nora, A., & Asker, E. H. (2000). Economic influences on persistence reconsidered. *Reworking the student departure puzzle*, 29-47.

Stewart, S., Lim, D. H., JoHyun, K. (2015). Factors influencing college persistence for first-time students. *Journal of Developmental Education, 38*(3).

Therriault, S. B., & Krivoshey, A. (2014). *College persistence indicators research review.* Association for Institutional Research.

Tinto, V. (1987). *Leaving college: Rethinking the causes and cures of student attrition.* Chicago: The University of Chicago Press.

U.S. Department of Education, National Center for Education Statistics. (2014). *STEM*

attrition: College students' paths into and out of stem fields

US News and World Report Best Colleges Methodology Retrieved from: http://www.usnews.com/education/best-colleges/articles/how-us-news-calculated-the-rankings?page=3

Walsh, W. B. (1973). The theory of person-environment interaction: Implications for the college student. Iowa City, IA: American College Testing Program.

Weiss, G. M. (2004). Mining with rarity: a unifying framework. *ACM Sigkdd Explorations Newsletter*, *6*(1), 7-19.

Wu, D., Fletcher, K., & Olson, L. (2008). A study of college student attrition via probabilistic approach. *The Journal of Mathematical Sociology, 31*(1), 89-95. doi: 10.1080/ 00222500600561238

Appendix A

An FTFY student is the intersection of two specific student characteristics which are federally defined. IPEDS defines an undergraduate first-time student as:

> A student who has no prior postsecondary experience (except as noted below) attending any institution for the first time at the undergraduate level. This includes students enrolled in academic or occupational programs. It also includes students enrolled in the fall term who attended college for the first time in the prior summer term, and students who entered with advanced standing (college credits earned before graduation from high school) (IPEDS, 2016).

In addition to possessing the above qualities of a first-time student, an FTFY student must also possess the qualities of a first-year student, which IPEDS describes as:

> A student who has completed less than the equivalent of one full year of undergraduate work; that is, less than 30 semester hours (in a 120-hour degree program) or less than 900 contact hours. (IPEDS, 2016).

Appendix B

Variable	Description
ID	unique identifier
COHORT_WEEK3_FIRSTTIME	student cohort stamp
AGE_Q1EOT	student age
First_Gen	first-generation status
GENDER	student gender
Persist_Q5WK3	persistence indicator
ADMINISTRATOR_RATING	student admission rating
SORGPAT_GPA	high school GPA
TOTAL_ACCEPT_AMOUNT	total amount of finanical aid accepted
TOTAL_OFFER_AMOUNT	total amount of finanical aid offered
FIN_AID_TYPE_Grant	amount of grant aid
FIN_AID_TYPE_Loan	amount of loan aid
FIN_AID_TYPE_Scholarship	amount of scholarship aid
FIN_AID_TYPE_Work	amount of aid from student employment
GIFT_OR_SELF_HELP_Gift Aid	amount of gift aid
GIFT_OR_SELF_HELP_Self Help Aid	amount of self help aid
FUND_SOURCE_DESC_Departmental Funded Schl	amount of scholarship funded by department
FUND_SOURCE_DESC_Federal Aid	amount of federal aid
FUND_SOURCE_DESC_Private Funding	amount of private funding
FUND_SOURCE_DESC_State Funding	amount of state funding
FUND_SOURCE_DESC_UGrad Gift & Endowed	amount of undergraduate gift aid
FUND_SOURCE_DESC_Undergraduate Discount	amount of undergraduate discount
FED_FUND_ID_CWS	Amount of College Work-Study Funding
FED_FUND_ID_PELL	Amount of Pell funding
FED_FUND_ID_PERK	Amount of Perkins loan funding
FED_FUND_ID_PLUS	Amount of PLUS loan funding
FED_FUND_ID_SEOG	Amount of Federal Supplemental Educational Opportunity Grant funding
FED_FUND_ID_STFD	Amount of Stafford loan funding
DSF	Denver Scholarship Foundation indicator
No_Aid	Indicator for receipt of financial aid
IM_UNMET_NEED	Amount of unmet financial need
FM_APPLICATION_IND	Federal aid application submission indicator
MINOR_1_Q1EOT	Minor, end of first term
ATHLETE_Q1EOT	athlete indicator, end of first term

Activity_Count_Q1	count of activities during first term
GREEK_Q1EOT	Greek organization indicator, end of first term
Intramurals_Activ_Type_Q1	intramurals participation indicator, first term
LEP_Q1	learning effectiveness program participation indicator, first term
Honors Program_Q1	honors program participant, first term
Living & Learning Communities_Activ_Type_Q1	living and learning community participant indicator, first term
Social_Activ_Type_Q1	social activity participation indicator, first term
HOUSING_IND_Q1EOT	university housing indicator, end of first term
ADVISOR_COUNT_Q1	count of student advisors, first term
Active_Holds_Q1	count of active holds, end of first term
HoldsNew_Q1	count of new holds during first term
Active_REGISTRATION_Holds_Q1	count of active registration holds, end of first term
Active_APPLICATION_Holds_Q1	count of active application holds, end of first term
HOLD_COUNT_Q1	count of all holds, first term
GRADE_VALUE_FSEM	grade value in first-year seminar
W_Count_Q1	count of course withdrawals, first term
min_GradeVal_Q1	minimum grade value, first term
max_GradeVal_Q1	maximum grade value, first term
mean_GradeVal_Q1	mean grade value, first term
Fail_Count_Q1	count of failed courses, first term
Proportion_belowCminus_Q1	proportion of courses with grades below a C-, first term
CREDITS_ATTEMPTED_Q1	total credits attempted, first term
CREDITS_EARNED_Q1	total credits earned, first term
GPA_Q1	GPA, first term
REGISTERED_HRS_Q1WK3	total registered credit hourse, third week first term
REGISTERED_HRS_Q1EOT	total registered credit hourse, end of first term
TRANSFER_HRS_Q1EOT	total transfer credit hours, end of first term
CUMULATIVE_HRS_BOT_Q1WK3	total cummulative credit hours earned at institution, third week first term
Academic_Difficulty_Q1	count of academic difficulty report, first term
Academic_Misconduct_Q1	count of academic misconduct report, first term
Cares_Submissions_Q1	count of wellness concerns, first term
Cleanliness_Q1	count of cleanliness violation, first term
Death_Q1	count of concerns related to death, first term
Dishonesty_Q1	count of dishonesty violation, first term
Disorderly_Conduct_Q1	count of disorderly conduct violation, first term
Drugs_Alcohol_Q1	count of drug or alcohol violation, first term

Endangerment_Weapons_Provoke_Q1	count of violations related to endangerment, weapons or provocation, first term
Harassment_Q1	count of report, first term
Level1_Compliance_Q1	count of level1 compliance report, first term
Mental_Health_Q1	count of mental health report, first term
Missing_Q1	count of report, first term
Physcial_Health_Q1	count of physcial health report, first term
PropertyDamage_Theft_Q1	count of propertydamage theft report, first term
Sexual_EEO_Q1	count of sexual eeo report, first term
ACT Comp Conv	converted and combined ACT composite score
ACT Math Conv	converted and combined ACT mathematics score
ACT English Conv	converted and combined ACT verbal score
Cof	state opportunity fund recipient indicator
RACE_DESC	race description
ETHN_CDE_DESC	ethnicity description
Race_Ethn	student race/ethnicity
CITZ_CODE	student citizenship indicator
State	in-state, out-of-state, international student classification
STUDENT_CLASSIFICATION_DESC	student classification (e.g., freshman, sophomore)
BUILDING_DESC_Q1	residence building description for first term
COLLEGE_1_Q1EOT	primary college, end of first term
MAJOR_1_Q1EOT	primary major, end of first term
DEGREE_1_Q1EOT	primary degree, end of first term
PROGRAM_1_Q1EOT	primary program, end of first term
DEPARTMENT_1_Q1EOT	primary department, end of first term
Total_Conduct_Q1	total number of conduct offenses during first term

Appendix C

	Cross-Validation Dataset	Evaluation Dataset
Variable	Mean (SD) or Proportion	Mean (SD) or Proportion
Age	18.30 (0.55)	18.30 (0.51)
Admit Rating	5.07 (2.80)	4.52 (2.82)
HS GPA	3.70 (0.34)	3.73 (0.33)
Intramural Activities	0.20 (0.40)	0.33 (0.47)
Activities Total	0.74 (0.91)	0.78 (0.92)
New Holds	0.25 (0.62)	0.38 (0.78)
Seminar Grade	85.83 (8.77)	86.23 (8.53)
Withdrawal	17.10 (0.46)	0.16 (0.45)
Min Grade	77.92 (10.07)	78.86 (10.15)
Max Grade	88.38 (5.36)	88.35 (5.97)
Credits attempted	16.21 (1.22)	15.99 (1.39)
GPA Q1	3.32 (0.62)	3.37 (0.62)
Transfer Hours	10.54 (14.92)	11.83 (15.74)
Total Conduct Violations	0.80 (1.83)	0.81 (2.13)
ACT Composite Converted	27.15 (3.37)	27.41 (3.43)
Male	45.94%	44.18%
Attrition	13.30%	13.21%
No Aid	14.59%	12.76%
Social Activities	22.92%	22.91%
Athlete	6.02%	5.82%

Learning Disability Services	6.02%	7.16%
Honors Program	7.29%	7.46%
Living Learning Community	15.20%	6.49%
University Housing	95.38%	96.49%

Appendix D

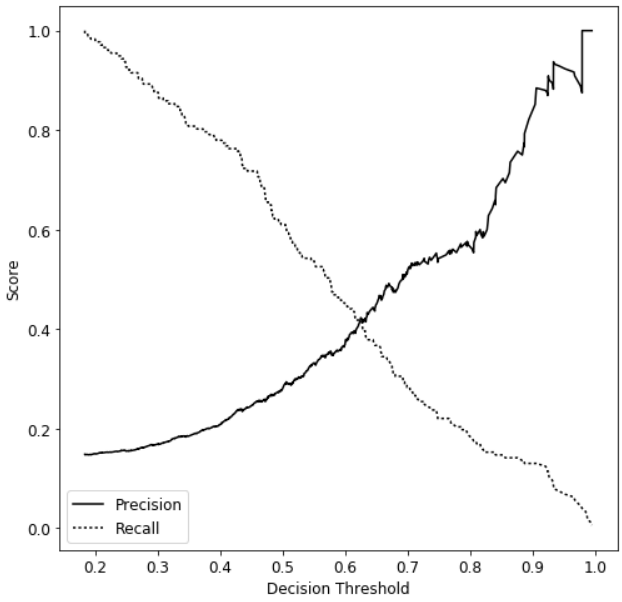


Figure D1. Precision and Recall as a Function of Decision Threshold: Logistic Regression

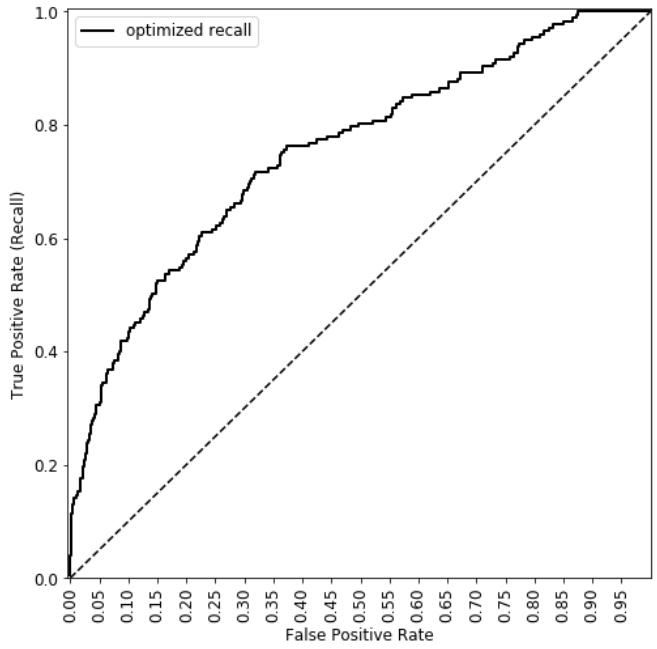


Figure D2. ROC Curve: Logistic Regression

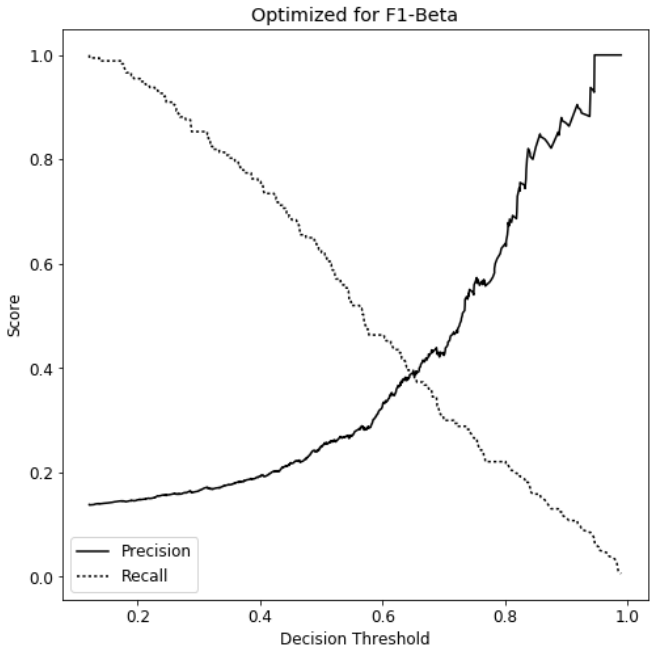


Figure D3. Precision and Recall as a Function of Decision threshold: F-Beta-Optimized MLP

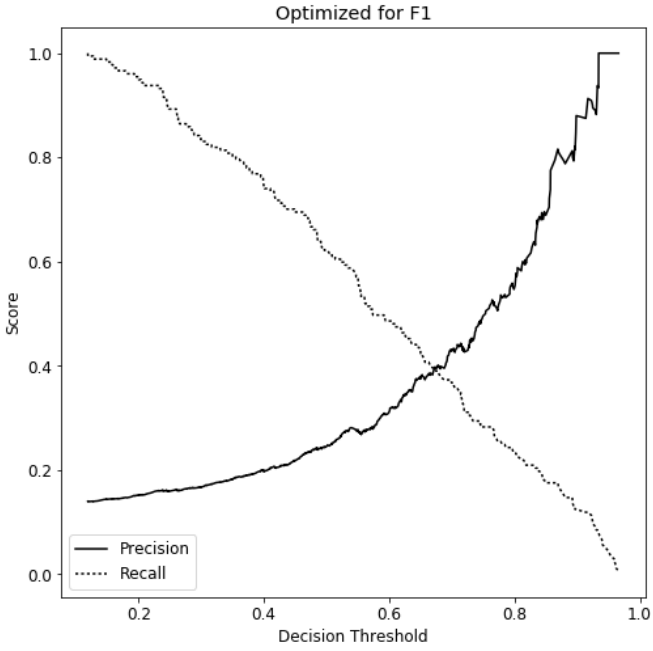


Figure D4. Precision and Recall as a Function of Decision threshold: F1-Optimized MLP

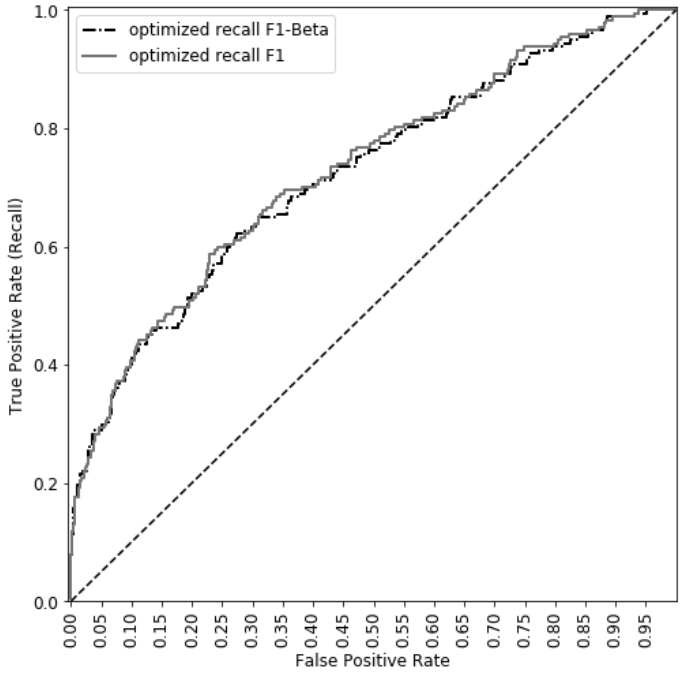


Figure D5. ROC Curve: MLP

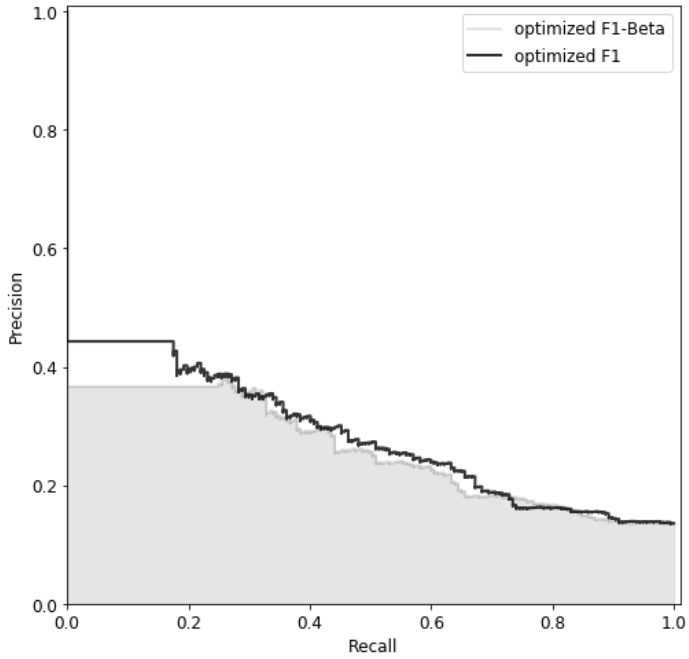


Figure D6. Precision-Recall Curve – Naïve Bayes

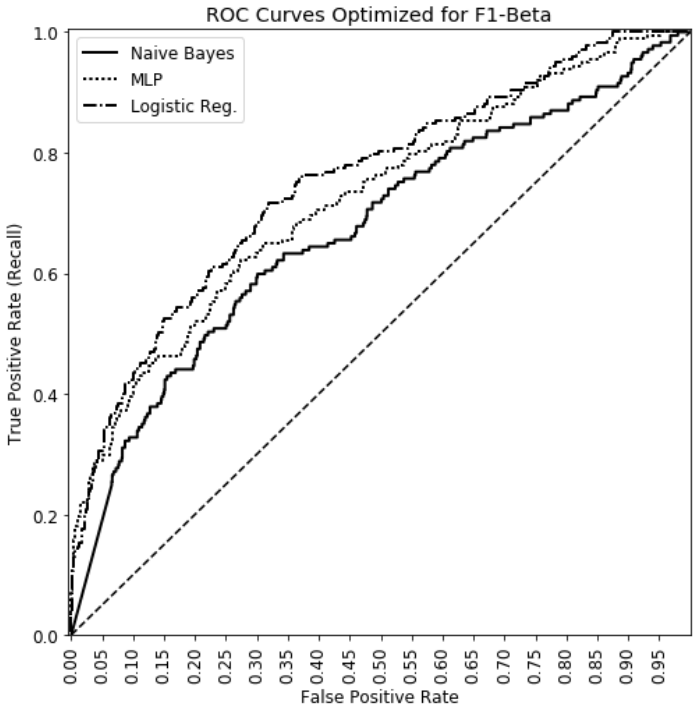


Figure D7. Overall ROC Curves Optimized for F-Beta

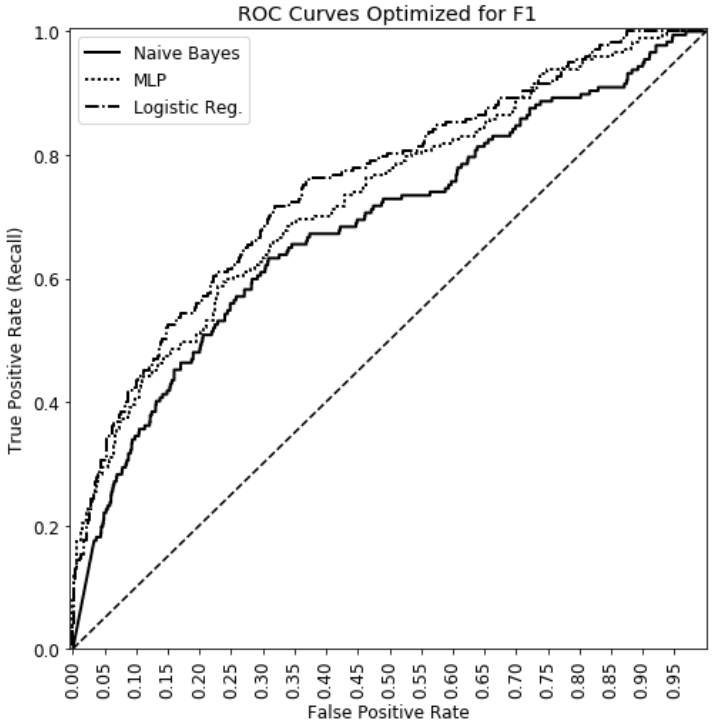


Figure D8. Overall ROC Curves Optimized for F1

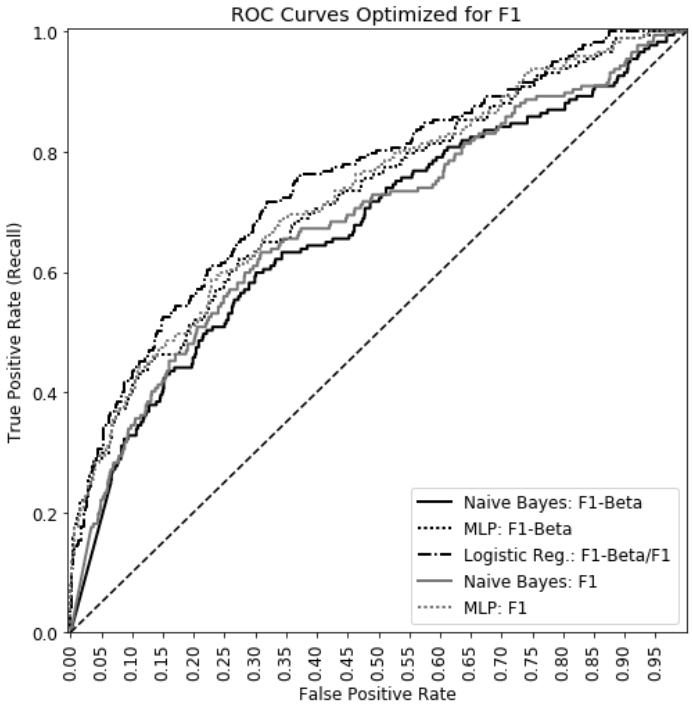
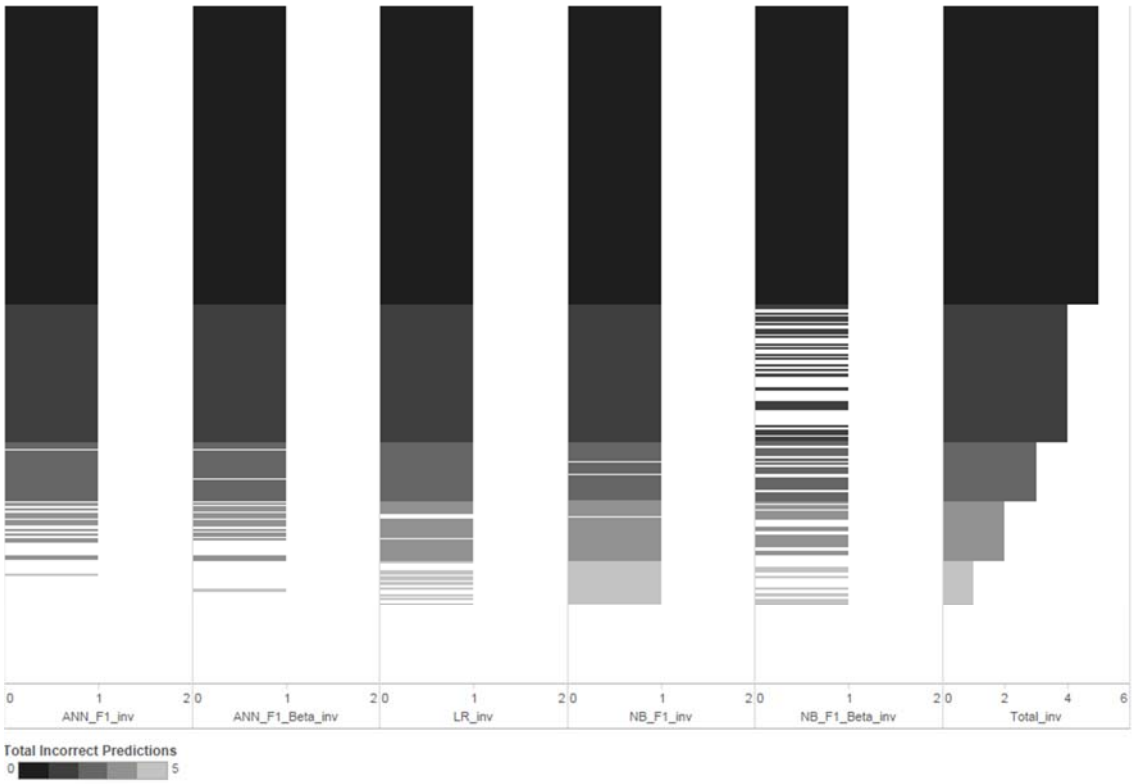


Figure D9. All ROC Curves

Appendix E

Correctly Classified Persistence Cases from Hold-Out by Model Type by Consensus



Appendix F

Feature Group	Feature Name
Top 50	Academic_Difficulty_Q1
Top 50	ACT Comp Conv
Top 50	ACT English Conv
Top 50	ACT Math Conv
Top 50	Active_APPLICATION_Holds_Q1
Top 50	Active_Holds_Q1
Top 50	Active_REGISTRATION_Holds_Q1
Top 50	Activity_Count_Q1
Top 50	ADMINISTRATOR_RATING
Top 50	ATHLETE_Q1EOT
Top 50	BUILDING_DESC_Q1_Centennial Towers North
Top 50	BUILDING_DESC_Q1_Centennial Towers South
Top 50	BUILDING_DESC_Q1_McFarlane Hall
Top 50	Cares_Submissions_Q1
Top 50	COLLEGE_1_Q1EOT_DC
Top 50	COLLEGE_1_Q1EOT_IS
Top 50	COLLEGE_1_Q1EOT_SS
Top 50	CREDITS_ATTEMPTED_Q1
Top 50	CREDITS_EARNED_Q1
Top 50	DEGREE_1_Q1EOT_BSAC
Top 50	DEGREE_1_Q1EOT_BSBA
Top 50	DEPARTMENT_1_Q1EOT_ANTH
Top 50	DEPARTMENT_1_Q1EOT_DCG
Top 50	DEPARTMENT_1_Q1EOT_GSIS
Top 50	DEPARTMENT_1_Q1EOT_HIST
Top 50	DEPARTMENT_1_Q1EOT_HRTM
Top 50	DEPARTMENT_1_Q1EOT_MGMT
Top 50	DEPARTMENT_1_Q1EOT_NMG
Top 50	DEPARTMENT_1_Q1EOT_PSYC
Top 50	DEPARTMENT_1_Q1EOT_RLGS
Top 50	Dishonesty_Q1
Top 50	Drugs_Alcohol_Q1
Top 50	DSF
Top 50	Endangerment_Weapons_Provoke_Q1
Top 50	ETHN_CDE_DESC_Not Hispanic or Latino
Top 50	Fail_Count_Q1
Top 50	FIN_AID_TYPE_Grant
Top 50	FIN_AID_TYPE_Scholarship

Top 50	FUND_SOURCE_DESC_Departmental Funded Schl
Top 50	FUND_SOURCE_DESC_Private Funding
Top 50	FUND_SOURCE_DESC_State Funding
Top 50	FUND_SOURCE_DESC_UGrad Gift & Endowed
Top 50	FUND_SOURCE_DESC_Undergraduate Discount
Top 50	GENDER
Top 50	GIFT_OR_SELF_HELP_Gift Aid
Top 50	GPA_Q1
Top 50	GRADE_VALUE_FSEM
Top 50	GREEK_Q1EOT
Top 50	HOLD_COUNT_Q1
Top 50	HoldsNew_Q1
Top 75	Honors Program_Q1
Top 75	HOUSING_IND_Q1EOT
Top 75	IM_UNMET_NEED
Top 75	Intramurals_Activ_Type_Q1
Top 75	LEP_Q1
Top 75	Living & Learning Communities_Activ_Type_Q1
Top 75	LLC:Pioneer Leadership Program_Q1
Top 75	MAJOR_1_Q1EOT_ANIG
Top 75	MAJOR_1_Q1EOT_ANTH
Top 75	MAJOR_1_Q1EOT_COMN
Top 75	MAJOR_1_Q1EOT_EBIO
Top 75	MAJOR_1_Q1EOT_EDPX
Top 75	MAJOR_1_Q1EOT_GBUS
Top 75	MAJOR_1_Q1EOT_HIST
Top 75	MAJOR_1_Q1EOT_HPM
Top 75	MAJOR_1_Q1EOT_INTS
Top 75	MAJOR_1_Q1EOT_PSYC
Top 75	MAJOR_1_Q1EOT_RLGS
Top 75	MAJOR_1_Q1EOT_UNBU
Top 75	MAJOR_1_Q1EOT_UNNS
Top 75	max_GradeVal_Q1
Top 75	mean_GradeVal_Q1
Top 75	Mental_Health_Q1
Top 75	min_GradeVal_Q1
Top 75	MINOR_1_Q1EOT
Top 100	Missing_Q1
Top 100	No_Aid
Top 100	PROGRAM_1_Q1EOT_BA-ECS
Top 100	PROGRAM_1_Q1EOT_BA-INTS

Top 100	PROGRAM_1_Q1EOT_BA-SOC SCI
Top 100	PROGRAM_1_Q1EOT_BSACC
Top 100	PROGRAM_1_Q1EOT_BSBA
Top 100	PropertyDamage_Theft_Q1
Top 100	Proportion_belowCminus_Q1
Top 100	RACE_DESC_Native Hawaiian or Other Pacific Islander
Top 100	Race_Ethn_Hispanic or Latino
Top 100	Race_Ethn_Native Hawaiian or Other Pacific Islander
Top 100	REGISTERED_HRS_Q1EOT
Top 100	REGISTERED_HRS_Q1WK3
Top 100	Sexual_EEO_Q1
Top 100	Social_Activ_Type_Q1
Top 100	SORGPAT_GPA
Top 100	State_out-of-state
Top 100	STUDENT_CLASSIFICATION_DESC_Q1_Sophomore
Top 100	TOTAL_ACCEPT_AMOUNT
Top 100	Total_Conduct_Q1
Top 100	TOTAL_CREDITS_Q1
Top 100	TOTAL_OFFER_AMOUNT
Top 100	TRANSFER_HRS_Q1EOT
Top 100	W_Count_Q1

Appendix G

Feature	Coefficient	SD	Coeficient x SD	ABS
FUND_SOURCE_DESC_Undergraduate Discount	-0.57	1	-0.57	0.57
TOTAL_ACCEPT_AMOUNT	-0.54	1	-0.54	0.54
ADMINISTRATOR_RATING	-0.27	1	-0.27	0.27
FUND_SOURCE_DESC_State Funding	0.25	1	0.25	0.25
mean_GradeVal_Q1	-0.21	1	-0.21	0.21
GIFT_OR_SELF_HELP_Gift Aid	-0.19	1	-0.19	0.19
CREDITS_EARNED_Q1	-0.19	1	-0.19	0.19
CREDITS_ATTEMPTED_Q1	-0.18	1	-0.18	0.18
Total_Conduct_Q1	-0.18	1	-0.18	0.18
TOTAL_CREDITS_Q1	0.18	1	0.18	0.18
TOTAL_OFFER_AMOUNT	-0.17	1	-0.17	0.17
HoldsNew_Q1	-0.16	1	-0.16	0.16
min_GradeVal_Q1	-0.16	1	-0.16	0.16
Activity_Count_Q1	-0.14	1	-0.14	0.14
FIN_AID_TYPE_Scholarship	0.14	1	0.14	0.14
Proportion_belowCminus_Q1	0.13	1	0.13	0.13
FUND_SOURCE_DESC_Departmental Funded Schl	0.13	1	0.13	0.13
Cares_Submissions_Q1	0.13	1	0.13	0.13
W_Count_Q1	-0.12	1	-0.12	0.12
Sexual_EEO_Q1	0.12	1	0.12	0.12
Endangerment_Weapons_Provoke_Q1	0.11	1	0.11	0.11
GPA_Q1	0.11	1	0.11	0.11
BUILDING_DESC_Q1_McFarlane Hall	-0.30	0.35	-0.11	0.11
Academic_Difficulty_Q1	0.10	1	0.10	0.10
Active_Holds_Q1	0.10	1	0.10	0.10
FUND_SOURCE_DESC_UGrad Gift & Endowed	0.09	1	0.09	0.09
STUDENT_CLASSIFICATION_DESC_ Q1_Sophomore	0.27	0.33	0.09	0.09
ACT Math Conv	0.09	1	0.09	0.09
ACT Comp Conv	0.09	1	0.09	0.09
Fail_Count_Q1	-0.08	1	-0.08	0.08
State_out-of-state	0.17	0.48	0.08	0.08
Dishonesty_Q1	0.08	1	0.08	0.08
REGISTERED_HRS_Q1EOT	0.08	1	0.08	0.08
Drugs_Alcohol_Q1	-0.08	1	-0.08	0.08
Mental_Health_Q1	0.07	1	0.07	0.07

HOLD_COUNT_Q1	0.07	1	0.07	0.07
COLLEGE_1_Q1EOT_DC	-0.16	0.41	-0.07	0.07
IM_UNMET_NEED	-0.06	1	-0.06	0.06
BUILDING_DESC_Q1_Centennial Towers South	0.19	0.28	0.05	0.05
ETHN_CDE_DESC_Not Hispanic or Latino	-0.17	0.31	-0.05	0.05
MAJOR_1_Q1EOT_UNBU	-0.23	0.23	-0.05	0.05
DEPARTMENT_1_Q1EOT_DCG	-0.23	0.23	-0.05	0.05
MAJOR_1_Q1EOT_PSYC	0.23	0.22	0.05	0.05
DEPARTMENT_1_Q1EOT_PSYC	0.23	0.22	0.05	0.05
Living & Learning Communities_Activ_Type_Q1	-0.20	0.25	-0.05	0.05
MAJOR_1_Q1EOT_UNNS	0.40	0.12	0.05	0.05
LLC:Pioneer Leadership Program_Q1	0.19	0.25	0.05	0.05
GREEK_Q1EOT	-0.11	0.41	-0.05	0.05
COLLEGE_1_Q1EOT_SS	-0.12	0.37	-0.05	0.05
REGISTERED_HRS_Q1WK3	-0.04	1	-0.04	0.04
Active_REGISTRATION_Holds_Q1	0.04	1	0.04	0.04
Social_Activ_Type_Q1	-0.10	0.42	-0.04	0.04
DEPARTMENT_1_Q1EOT_HRTM	-0.28	0.12	-0.03	0.03
DEPARTMENT_1_Q1EOT_HIST	0.35	0.08	0.03	0.03
TRANSFER_HRS_Q1EOT	-0.03	1	-0.03	0.03
No_Aid	-0.08	0.33	-0.03	0.03
Honors Program_Q1	-0.09	0.26	-0.02	0.02
max_GradeVal_Q1	0.02	1	0.02	0.02
SORGPAT_GPA	-0.02	1	-0.02	0.02
LEP_Q1	-0.07	0.26	-0.02	0.02
DEGREE_1_Q1EOT_BSAC	-0.14	0.12	-0.02	0.02
PROGRAM_1_Q1EOT_BSACC	-0.14	0.12	-0.02	0.02
GRADE_VALUE_FSEM	-0.02	1	-0.02	0.02
MAJOR_1_Q1EOT_ANTH	0.27	0.05	0.01	0.01
DEPARTMENT_1_Q1EOT_ANTH	0.27	0.05	0.01	0.01
HOUSING_IND_Q1EOT	0.06	0.18	0.01	0.01
DEGREE_1_Q1EOT_BSBA	-0.02	0.40	-0.01	0.01
PROGRAM_1_Q1EOT_BSBA	-0.02	0.40	-0.01	0.01
DEPARTMENT_1_Q1EOT_MGMT	-0.03	0.22	-0.01	0.01
Intramurals_Activ_Type_Q1	0.01	0.47	0.01	0.01
Active_APPLICATION_Holds_Q1	0	1	0	0
Missing_Q1	0.06	0	0	0
RACE_DESC_Native Hawaiian or Other Pacific Islander	0.32	0	0	0
MAJOR_1_Q1EOT_ANIG	0.19	0	0	0

MAJOR_1_Q1EOT_GBUS	-0.54	0	0	0

Appendix H

Correct model predictions	0	1	2	3	4	5
n	36	16	12	22	23	68
				M (SD)		
AGE_Q1EOT	18.28 (0.51)	18.31 (0.46)	18.5 (0.5)	18.27 (0.45)	18.48 (0.58)	18.28 (0.56)
First_Gen	0.22 (0.42)	0.19 (0.39)	0.25 (0.43)	0.18 (0.39)	0.09 (0.28)	0.13 (0.34)
GENDER	0.33 (0.47)	0.31 (0.46)	0.08 (0.28)	0.59 (0.49)	0.48 (0.5)	0.65 (0.48)
ADMINISTRATOR_RATING	3.81 (2.28)	4.56 (2.98)	5.33 (2.9)	4.91 (2.33)	5.57 (2.39)	6.87 (2.41)
SORGPAT_GPA	3.79 (0.26)	3.76 (0.27)	3.64 (0.34)	3.64 (0.37)	3.66 (0.29)	3.43 (0.38)
DSF	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0.01 (0.12)
No_Aid	0 (0)	0.06 (0.24)	0.25 (0.43)	0.14 (0.34)	0.13 (0.34)	0.41 (0.49)
FM_APPLICATION_IND	0.61 (0.49)	0.75 (0.43)	0.75 (0.43)	0.64 (0.48)	0.74 (0.44)	0.62 (0.49)
MINOR_1_Q1EOT	0.11 (0.31)	0.06 (0.24)	0.17 (0.37)	0 (0)	0 (0)	0.07 (0.26)
ATHLETE_Q1EOT	0.14 (0.35)	0.06 (0.24)	0 (0)	0.09 (0.29)	0.04 (0.2)	0.01 (0.12)
Activity_Count_Q1	1.08 (0.95)	1.06 (0.83)	0.83 (0.8)	0.23 (0.42)	0.52 (0.77)	0.4 (0.71)
GREEK_Q1EOT	0.31 (0.46)	0.44 (0.5)	0.42 (0.49)	0 (0)	0.04 (0.2)	0.09 (0.28)
Intramurals_Activ_Type_Q1	0.47 (0.5)	0.5 (0.5)	0.33 (0.47)	0.23 (0.42)	0.35 (0.48)	0.24 (0.42)
LEP_Q1	0.03 (0.16)	0 (0)	0.17 (0.37)	0.14 (0.34)	0.09 (0.28)	0.18 (0.38)
Honors Program_Q1	0.11 (0.31)	0.06 (0.24)	0 (0)	0 (0)	0 (0)	0 (0)
Living & Learning Communities_Activ_Type_Q1	0.08 (0.28)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
LLC:Pioneer Leadership Program_Q1	0.08 (0.28)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
Social_Activ_Type_Q1	0.31 (0.46)	0.44 (0.5)	0.5 (0.5)	0 (0)	0.04 (0.2)	0.09 (0.28)
HOUSING_IND_Q1EOT	1 (0)	1 (0)	1 (0)	0.95 (0.21)	0.96 (0.2)	0.97 (0.17)
ADVISOR_COUNT_Q1	2.64 (0.95)	2.81 (0.81)	2.83 (1.07)	2.32 (0.63)	2.48 (0.77)	2.57 (0.75)
Active_Holds_Q1	0 (0)	0.06 (0.24)	0.08 (0.28)	0 (0)	0 (0)	0.1 (0.3)
HoldsNew_Q1	0.33 (0.67)	0.25 (0.56)	0.75 (0.83)	0.5 (0.58)	0.22 (0.51)	0.63 (0.78)
Active_REGISTRATION_Holds_Q1	0 (0)	0.06 (0.24)	0 (0)	0 (0)	0 (0)	0.07 (0.26)

Active_APPLICATION_Holds_Q1	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0.01 (0.12)
HOLD_COUNT_Q1	0.56 (0.68)	0.19 (0.39)	0.75 (0.6)	0.86 (0.92)	0.87 (0.68)	1.78 (1.64)
GRADE_VALUE_FSEM	87.78 (4.26)	85.56 (5.51)	86.58 (3.66)	88.14 (3.4)	86.22 (4.56)	73.72 (20.06)
W_Count_Q1	0.06 (0.23)	0.38 (0.6)	0.5 (0.65)	0.18 (0.49)	0.09 (0.28)	0.59 (0.94)
min_GradeVal_Q1	81.64 (5.31)	80.75 (7.5)	77.25 (12.08)	79.14 (6.55)	78.43 (4.27)	63.15 (16.48)
max_GradeVal_Q1	89.64 (1.16)	88.88 (2.6)	88.17 (3.05)	89.18 (2.27)	89.48 (1.06)	77.26 (17.64)
mean_GradeVal_Q1	86.85 (2.57)	85.35 (4.37)	83.79 (6.59)	85.16 (3.68)	84.45 (2.25)	70.71 (16.43)
Fail_Count_Q1	0 (0)	0 (0)	0.08 (0.28)	0 (0)	0 (0)	0.54 (1.08)
Proportion_belowCminus_Q1	0 (0)	0.02 (0.06)	0.06 (0.15)	0 (0)	0.01 (0.03)	0.38 (0.36)
CREDITS_ATTEMPTED_Q1	16.03 (1.07)	16 (1.22)	16.08 (1.44)	15.68 (1.18)	16.09 (0.65)	14.94 (1.98)
CREDITS_EARNED_Q1	15.81 (1.41)	14.56 (2.5)	14.42 (3.2)	15.14 (1.91)	15.74 (1.33)	10.96 (5.04)
GPA_Q1	3.59 (0.3)	3.44 (0.45)	3.28 (0.63)	3.42 (0.38)	3.34 (0.24)	2.16 (1.13)
REGISTERED_HRS_Q1WK3	16.03 (1.07)	15.5 (1.8)	15.75 (1.83)	15.5 (1.41)	16.09 (0.65)	14.68 (2)
REGISTERED_HRS_Q1EOT	15.81 (1.41)	14.56 (2.5)	14.67 (2.87)	15.14 (1.91)	15.74 (1.33)	13.34 (2.82)
TRANSFER_HRS_Q1EOT	13.08 (14.18)	12.81 (17.99)	10.46 (11.75)	7.23 (8.8)	7.11 (11.03)	5.79 (14.58)
CUMULATIVE_DU_HRS_BOT_Q1WK3	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
TOTAL_CREDITS_Q1	15.81 (1.41)	14.56 (2.5)	14.67 (2.87)	15.14 (1.91)	15.74 (1.33)	13.34 (2.82)
Academic_Difficulty_Q1	0 (0)	0 (0)	0 (0)	0 (0)	0.04 (0.2)	0.22 (0.64)
Academic_Misconduct_Q1	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0.03 (0.17)
Cares_Submissions_Q1	0.08 (0.36)	0.06 (0.24)	0.42 (0.76)	0.55 (1.16)	0.3 (0.86)	0.57 (1.28)
Cleanliness_Q1	0 (0)	0 (0)	0 (0)	0.09 (0.29)	0.04 (0.2)	0.06 (0.29)
Drugs_Alcohol_Q1	0.33 (1.03)	0 (0)	0.58 (1.66)	0.59 (1.78)	0.61 (1.71)	0.91 (1.95)
Endangerment_Weapons_Provoke_Q1	0 (0)	0 (0)	0.08 (0.28)	0.09 (0.29)	0.04 (0.2)	0.06 (0.24)
Harassment_Q1	0 (0)	0 (0)	0 (0)	0 (0)	0.04 (0.2)	0.01 (0.12)
Level1_Compliance_Q1	0.06 (0.23)	0 (0)	0.08 (0.28)	0 (0)	0.04 (0.2)	0.13 (0.34)
Mental_Health_Q1	0 (0)	0.06 (0.24)	0.08 (0.28)	0.36 (0.93)	0 (0)	0.21 (0.7)
Physcial_Health_Q1	0.06 (0.33)	0 (0)	0 (0)	0.05 (0.21)	0 (0)	0.03 (0.17)
PropertyDamage_Theft_Q1	0.03 (0.16)	0 (0)	0.08 (0.28)	0 (0)	0 (0)	0.04 (0.21)

Sexual_EEO_Q1	0 (0)	0 (0)	0 (0)	0 (0)	0.13 (0.61)	0 (0)
ACT Comp Conv	27.92 (3.29)	27.56 (3.06)	25.92 (3.8)	27.36 (3.13)	27.43 (3.6)	26.54 (3.09)
ACT Math Conv	26.97 (3.81)	26.56 (2.65)	24.17 (3.58)	25.36 (3.72)	25.91 (4.86)	25.59 (3.36)
ACT English Conv	28.25 (5.51)	29.13 (3.89)	25.83 (5.11)	27.73 (3.93)	27.7 (5)	25.69 (5.23)
Total_Conduct_Q1	0.56 (1.5)	0.13 (0.48)	1.33 (3.01)	1.73 (2.91)	1.26 (2.47)	2.41 (4.09)
TOTAL_ACCEPT_AMOUNT	35800.89 (17368.65)	35549.96 (21130.79)	30473.58 (23674.98)	23812.53 (19593.27)	20220.1 (18268.83)	14124.47 (16630.8)
TOTAL_OFFER_AMOUNT	35780.11 (17341)	35549.96 (21130.79)	30473.58 (23674.98)	23812.53 (19593.27)	20220.1 (18268.83)	14124.47 (16630.8)
FIN_AID_TYPE_Grant	6435.36 (9555.47)	8006.19 (10151.24)	5510.83 (8050.42)	6258.86 (9491.71)	2371.83 (4483.2)	3836.84 (7488.32)
FIN_AID_TYPE_Loan	6555.28 (11103.47)	9333 (14791.37)	8943.33 (14090.74)	4333.95 (8298.4)	7242.78 (14724.87)	3791.25 (7772.68)
FIN_AID_TYPE_Scholarship	22607.47 (10187.98)	18100.25 (10641.05)	15727.75 (10194.42)	13046.68 (7706.06)	10488.87 (7983.98)	6131.1 (7492.7)
FIN_AID_TYPE_Work	202.78 (673.09)	110.52 (428.04)	291.67 (967.35)	173.03 (541.59)	116.62 (509.49)	365.28 (853.77)
GIFT_OR_SELF_HELP_Gift Aid	29042.83 (14860.07)	26106.44 (15388.18)	21238.58 (14919.23)	19305.55 (13981.66)	12860.7 (9164.59)	9967.94 (12044.97)
GIFT_OR_SELF_HELP_Self Help Aid	6758.06 (11092.09)	9443.52 (14809.79)	9235 (13937.99)	4506.98 (8587.79)	7359.4 (14716.92)	4156.53 (7851.63)
FUND_SOURCE_DESC_Departmental Funded Schl	443.69 (2310.35)	2088.06 (5756.24)	0 (0)	1982.59 (6512.61)	107.92 (506.21)	245.59 (1695.16)
FUND_SOURCE_DESC_Federal Aid	5753.69 (8684.74)	10372.38 (14958.4)	6458.75 (9223.54)	5172.36 (9163.11)	5413.43 (11668.25)	4019.79 (7479.69)
FUND_SOURCE_DESC_Private Funding	3285.28 (8725.7)	3281.56 (9318.02)	3570.42 (10968.32)	397.27 (1270.57)	2973.91 (9258.47)	781.57 (4025.63)
FUND_SOURCE_DESC_State Funding	180.56 (647.14)	298.02 (1154.22)	0 (0)	305.3 (1084.94)	0 (0)	14.71 (120.37)
FUND_SOURCE_DESC_UGrad Gift & Endowed	763.89 (3512.85)	375 (1452.37)	0 (0)	171.91 (787.79)	0 (0)	0 (0)
FUND_SOURCE_DESC_Undergraduate Discount	25373.78 (9729.41)	19134.94 (9607.71)	20444.42 (13984.81)	15783.09 (11795.96)	11724.83 (8549.34)	9062.81 (10835.32)
FED_FUND_ID_CWS	133.33 (550.25)	0 (0)	291.67 (967.35)	140.45 (529.33)	108.7 (509.83)	328.51 (813.96)
FED_FUND_ID_PELL	822.78 (1769.11)	726.88 (1923.13)	544.17 (1233.8)	595.23 (1678.48)	644.57 (1678.92)	444.15 (1423.05)
FED_FUND_ID_PERK	222.22 (916.25)	343.75 (1331.34)	458.33 (1520.12)	326.18 (1038.93)	0 (0)	295.29 (978.45)
FED_FUND_ID_PLUS	2029.03 (6714.84)	6355.63 (13261.89)	3039.58 (6954.45)	2087.73 (6675.19)	2652.91 (10595.75)	1600 (5660.06)
FED_FUND_ID_SEOG	166.67 (687.18)	312.5 (845.48)	0 (0)	272.73 (862.44)	86.96 (407.86)	117.65 (556.5)
FED_FUND_ID_STFD	2379.67 (2579.8)	2633.63 (2668.09)	2125 (2566.82)	1750.05 (2439.59)	1920.3 (2446.46)	1234.19 (1830.77)
IM_UNMET_NEED	-9207.57 (17327.2)	-12186.89 (17428.11)	-5011.83 (10784.94)	-3460.43 (10392.34)	694.9 (19211.35)	2188.44 (13239.32)
RACE_DESC_Asian	0.06 (0.23)	0.13 (0.33)	0 (0)	0.09 (0.29)	0 (0)	0.01 (0.12)

RACE_DESC_Black or African American	0 (0)	0 (0)	0 (0)	0.05 (0.21)	0 (0)	0.03 (0.17)
RACE_DESC_Multiple (two or more races)	0.06 (0.23)	0 (0)	0 (0)	0 (0)	0.04 (0.2)	0 (0)
RACE_DESC_Native Hawaiian or Other Pacific Islander	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
RACE_DESC_Unknown	0.06 (0.23)	0 (0)	0.08 (0.28)	0.09 (0.29)	0.04 (0.2)	0.03 (0.17)
RACE_DESC_White	0.83 (0.37)	0.88 (0.33)	0.92 (0.28)	0.77 (0.42)	0.91 (0.28)	0.93 (0.26)
ETHN_CDE_DESC_None	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
ETHN_CDE_DESC_Not Hispanic or Latino	0.94 (0.23)	0.94 (0.24)	0.67 (0.47)	0.86 (0.34)	0.91 (0.28)	0.88 (0.32)
Race_Ethn_Asian	0.06 (0.23)	0.13 (0.33)	0 (0)	0.05 (0.21)	0 (0)	0 (0)
Race_Ethn_Black or African American	0 (0)	0 (0)	0 (0)	0.05 (0.21)	0 (0)	0.01 (0.12)
Race_Ethn_Hispanic or Latino	0.06 (0.23)	0.06 (0.24)	0.33 (0.47)	0.14 (0.34)	0.09 (0.28)	0.12 (0.32)
Race_Ethn_International	0.03 (0.16)	0 (0)	0 (0)	0.05 (0.21)	0 (0)	0.01 (0.12)
Race_Ethn_Multiple (two or more races)	0.06 (0.23)	0 (0)	0 (0)	0 (0)	0.04 (0.2)	0 (0)
Race_Ethn_Native Hawaiian or Other Pacific Islander	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
Race_Ethn_Unknown	0.03 (0.16)	0 (0)	0 (0)	0.05 (0.21)	0 (0)	0.03 (0.17)
Race_Ethn_White	0.78 (0.42)	0.81 (0.39)	0.67 (0.47)	0.68 (0.47)	0.87 (0.34)	0.82 (0.38)
CITZ_CODE_Y	0.97 (0.16)	1 (0)	1 (0)	0.95 (0.21)	1 (0)	0.99 (0.12)
State_International	0.03 (0.16)	0 (0)	0 (0)	0.05 (0.21)	0 (0)	0.01 (0.12)
State_out-of-state	0.72 (0.45)	0.63 (0.48)	0.75 (0.43)	0.55 (0.5)	0.96 (0.2)	0.74 (0.44)
STUDENT_CLASSIFICATION_DESC_Q1_Junior	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0.01 (0.12)
STUDENT_CLASSIFICATION_DESC_Q1_Senior	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
STUDENT_CLASSIFICATION_DESC_Q1_Sophomore	0.19 (0.4)	0.19 (0.39)	0.17 (0.37)	0.09 (0.29)	0.09 (0.28)	0.03 (0.17)
BUILDING_DESC_Q1_Centennial Halls North	0.11 (0.31)	0.19 (0.39)	0.33 (0.47)	0.32 (0.47)	0.35 (0.48)	0.34 (0.47)
BUILDING_DESC_Q1_Centennial Halls South	0.39 (0.49)	0.38 (0.48)	0 (0)	0.18 (0.39)	0.17 (0.38)	0.16 (0.37)
BUILDING_DESC_Q1_Centennial Towers North	0.19 (0.4)	0.13 (0.33)	0.17 (0.37)	0.18 (0.39)	0.13 (0.34)	0.15 (0.35)
BUILDING_DESC_Q1_Centennial Towers South	0.03 (0.16)	0.19 (0.39)	0.08 (0.28)	0 (0)	0.17 (0.38)	0.1 (0.3)

BUILDING_DESC_Q1_Johnson Hall	0.14 (0.35)	0 (0)	0.17 (0.37)	0.18 (0.39)	0.09 (0.28)	0.19 (0.39)
BUILDING_DESC_Q1_McFarlane Hall	0.14 (0.35)	0.13 (0.33)	0.25 (0.43)	0.09 (0.29)	0.04 (0.2)	0.03 (0.17)
COLLEGE_1_Q1EOT_DC	0.42 (0.49)	0.13 (0.33)	0.17 (0.37)	0.41 (0.49)	0 (0)	0.04 (0.21)
COLLEGE_1_Q1EOT_EN	0.06 (0.23)	0.06 (0.24)	0.08 (0.28)	0.14 (0.34)	0.09 (0.28)	0.18 (0.38)
COLLEGE_1_Q1EOT_IS	0 (0)	0 (0)	0 (0)	0.09 (0.29)	0.09 (0.28)	0.01 (0.12)
COLLEGE_1_Q1EOT_NM	0.22 (0.42)	0.44 (0.5)	0.33 (0.47)	0.09 (0.29)	0.22 (0.41)	0.19 (0.39)
COLLEGE_1_Q1EOT_SS	0.14 (0.35)	0.13 (0.33)	0.17 (0.37)	0 (0)	0.26 (0.44)	0.12 (0.32)
COLLEGE_1_Q1EOT_UG	0.11 (0.31)	0.19 (0.39)	0.08 (0.28)	0.27 (0.45)	0.22 (0.41)	0.37 (0.48)
DEPARTMENT_1_Q1EOT_UGG	0.11 (0.31)	0.19 (0.39)	0.08 (0.28)	0.27 (0.45)	0.22 (0.41)	0.37 (0.48)