Predicting underperformance from students in upper level engineering courses

William J. Darlington
wjd8283@rit.edu

Follow this and additional works at: http://scholarworks.rit.edu/theses

Recommended Citation

This Thesis is brought to you for free and open access by the Thesis/Dissertation Collections at RIT Scholar Works. It has been accepted for inclusion in Theses by an authorized administrator of RIT Scholar Works. For more information, please contact ritsscholarworks@rit.edu.
Predicting underperformance from students in upper level engineering courses

William J. Darlington

A Thesis Submitted in Partial Fulfillment of the Requirements for the Degree of Master of Science in Industrial Engineering

Department of Industrial and Systems Engineering
Kate Gleason College of Engineering

Rochester Institute of Technology
Rochester, New York
August 9, 2017
Committee Approval:

Rachel T. Silvestrini, Ph.D.
Kate Gleason Chair
Associate Professor
Thesis Advisor

Scott E. Grasman, Ph.D.
Department Head of Industrial and Systems Engineering
Professor
Committee Member
0 Abstract

Recent research in academic analytics has focused on predicting student performance within, and sometimes across courses for the purpose of informing early interventions. While such an endeavor has obvious merit, modern constructivist learning theory expresses an importance on more individualized support for students. In keeping with this theory, this research describes the development of a model that predicts student performance within a course, relative to their past academic performance. This study is done using the minimum sources of data possible while still developing an accurate model. Useful logistic models using data from the institution’s student information system, learning management system, and grade books some useful findings are developed. While each source of data was able to predict student success independently, the most accurate model contained data from both the grade book and student information system. These models were able to accurately identify students on track to underperform relative to their own cumulative grade point averages within the first seven weeks of a course, aligning with the studied institution’s existing requirements for a manual early intervention system.
1 Introduction

This thesis presents research regarding the use of current technologies used in academia for the development of a model that will identify students that are at risk of underperforming. The goal is to work towards a predictive model that can be used to identify students who might benefit from an intervention in regards to their overall performance in a course. In the future, such a model could be built into a system that automatically generates results and sends students automated messages, warning them of their risk of underperformance. Such a system is desirable for two reasons: it could both increase instructor participation in existing student feedback systems and it would present students with information to better make decisions regarding their academic careers.

Technology and academia have a long history of interdependence. This relationship has never been more obvious than with the widespread adoption of Learning Management Systems (LMSs), the increasing commonality of online courses, and the incorporation of eLearning in curriculum [1]. A LMS is a software suite that automates the administration and tracking of educational content, such as lectures, trainings, and assessments [2]. This new landscape for learning has generated a previously unfathomable amount of data regarding the ways in which students learn and the methods professors use to teach. The relatively new field of academic analytics is now more feasible than ever and gaining considerable interest, which is largely due to the availability of data.

Academic analytics can be defined as: using data about students, teachers, and learning environments to help compare courses and institutions, allowing for better decisions regarding policies ranging from acceptance criteria to course structures [3]. Within academic analytics, there exists a focus area known as learning analytics. According to the first International Conference for Learning Analytics and Knowledge, learning analytics is “the measurement, collection, analysis, and reporting of data about learners and their contexts, for the purpose of understanding and optimizing learning and the environments in which it occurs” [4].

Learning analytics have been used for quite some time, even before the advent of the online classroom. As early as the 1970s, statisticians were compiling mathematical models to help generate actionable intelligence for academic institutions. One of the more frequently
referenced authors, Vincent Tinto, modeled voluntary attrition rates among college students at Syracuse University. His main sources of data were from a Student Information System (SIS) (a database populated with students' personal and scholastic information), and surveys mailed to students to be completed on a voluntary basis [5]. Today, the bulk of data is collected by non-invasive sensors, meaning it is done in ways that do not interfere with, or require the participation of students. These sources, including records stored in LMS logs, have been accepted as less biased, as well as more thorough, and thus are currently preferred over direct observation and self-reporting by students as a method of data collection [6].

With these new sources of data, some universities have set out to develop models for student performance in specific programs. Current research in learning analytics is partially driven by the potential to improve student performance within classrooms. At Purdue University, an early alert system known as Course Signals has been developed and is being introduced in select pilot classes. This system uses proprietary algorithms to predict course outcomes for individual students in real time. Within those courses' LMSs, there is an indicator that reassures successful students of their progress and warns students that are at risk of failing, allowing for corrective action to be made [7]. Within the pilot classes, Purdue has seen a decrease in dropout rates; first year retention among the 2007 test group was 96.71% in the subset that utilized Course Signals, as opposed to 83.44% in the subset that did not. This is an undeniably desirable result that may not be easily replicated due to the lack of disclosure regarding the methodology [8]. This leaves a need for research regarding the development of a transparent predictive model of student performance, which may to encourage widespread adaptation and adoption of these types of models.

This need is highlighted by the existence of alternative programs that allow professors to manually mark students as being at risk of failure and notify the student, the student's advisors, and other professors. It is well documented that providing feedback of this nature helps student's feel empowered, which is a key attribute for an effective learning environment [9]. The institution being examined in this study currently utilizes such a system (Starfish EARLY ALERTS). As it exists at the time of this study, the institution’s EARLY ALERTS system allowed instructors to flag students as “at risk”. Students can be flagged based on low homework scores,
low exam scores, or attendance issues. When flagged the student is notified, along with
instructors for their other courses and faculty in their department.

The institution examined in this study requires for faculty participation, it is unenforced
so its use varies by department, professor, and even course. Additionally, this is a time-
consuming process for professors, and the accuracy of the predictions of failure is dependent on
their experience teaching and the extent of their interactions with each individual student. An
automated alert system would serve to both increase the quality of the service and decrease the
number of man-hours it requires.

While it is important to identify a student who is possibly at risk of failing a course, there
are other students not at risk of failing who might also benefit from feedback. According to
Hattie and Timperley, feedback is needed to explain the discrepancy between desired
performance and actual performance [9]. Consider the “straight A” student that receives a C in a
course. While the student did not fail, they will likely be unsatisfied with their grade relative to
their past performance. This example illustrates the need to broaden the qualifications for
students in need of feedback. Keeping with the understanding of feedback’s purpose from Hattie
and Timperley, this study will define a student’s need for feedback based on the discrepancy
between a student’s past performance and their performance in the course being modeled. This
method still requires that the students’ course performance be modeled which has its own set of
challenges.

One of the challenges in developing a predictive model for course performance is
selecting the correct features, or variables, to be included in the model. Past studies provide
some insight into what variables may be statistically significant factors for predicting student
grades. In many models, a student's cumulative grade point average (CGPA) is used as a factor,
as it is an accessible measure of a student's typical performance in the classroom [10–13]. In
addition to a measure of average academic performance, some researchers have attempted to
include estimates of subject matter knowledge.

There are a few different approaches to evaluating student’s subject matter knowledge.
One approach is to simply administer an examination at the start of a course that tests the
student's understanding of topics fundamental to the course. An alternative, non-invasive
method of capturing subject matter knowledge is to examine the student's performance in prerequisite courses. This can simply include the grade earned in a prerequisite class or it may go further to include the time elapsed between the prerequisite and current course [14].

Additional variables that have proven useful for predicting course performance are gender, age, and seating location. Unfortunately, there remains other information that is potentially relevant to student performance that cannot be directly measured, such as student interest and learning style. For these intrinsic factors, surveys remain a primary means of data collection due to the inability to effectively capture the information otherwise, but great care must be put into their design and administration [15]. Despite abiding by the rigorous requirements of administering one of these surveys, they remain subject to the biases inherent to self-reporting. For this reason, data available in LMSs has been used to supplement survey data, and sometimes replace it all together [6].

Using data typically captured by an LMS, researchers have quantified two student behaviors that play a role in success in classes: procrastination and engagement. In a 2015 paper, You [16] utilizes absences from class and late submissions of assignments as behavioral indicators of procrastination, and shows a correlation with their occurrence and lower course grades [16]. Junco and Clem use frequency of highlighting in conjunction with average duration of access and regularity of access of digital textbooks to quantify engagement, which shows a positive correlation with course outcomes [6]. Additionally, there has been significant research into methods of evaluating the various learning management systems that are available to colleges and universities. Consideration has been given to administrative ease, cost, user interface, and even the individual tools available within each system [17]. At an individual course level, Wilson and Fowler found that organizing classes to match an active-learning pedagogy resulted in students practicing more “deep learning” techniques [18]. This is an important finding because it shows that the way a professor organizes a class can impact a student's depth of understanding of a topic, and by extension their performance in that course, and thus, may be a possible predictor of performance.

This study aims to provide a framework for instructors to identify students who will earn a course grade that is uncharacteristically low or undesirable grade, as determined by a negative
impact on their CGPA. This framework involves the development of predictive models constructed utilizing survey data, historical data on academic performance, in-course assessment results, and LMS usage statistics. Models were generated using data restricted by the date it became available to investigate how the models, and predictions, change as a course progresses. It is expected that the accuracy of these predictions will improve as the course progresses, as demonstrated by Olama in a 2014 study of massive open online courses (MOOCs)[19]. While accuracy is important, the benefits of predicting course outcome lessen cover the timeframe of the course as there is both less time for corrective action to be taken and typically fewer degrees of freedom in the course grade. For this study, focus will be given to the weeks in which the institution being studied currently encourages instructors to identify at risk students.

This paper is divided into six sections: an introduction, literature review, data and limitations, methodology, results, and discussion. The literature review presents past research in the field of learning analytics, highlighting the gaps between and shortcomings of the studies. The Data and Limitations section explains how the data was collected for the study as well as the difficulties and limitations associated with them. The methodology is broken into two subsections, the first explaining how variables were transformed and used to generate new variables, and the second to explain how models were generated. The results section presents and explains the models created by following the methodology, and the discussion section interprets those results, drawing conclusions about students and suggesting next steps to take this research further.

2 Literature Review

This section summarizes selected past research in learning and academic analytics. The literature covered has provided guidance for this study in terms of what strategies have shown promise when predicting course outcomes, as well as showing challenges and shortcomings of past student performance research.

As part of the rise in eLearning tool, there has been a significant increase in the use of digital textbooks in higher education. In a 2015 study, students in 11 different courses were provided with digital copies of required textbooks on an online platform called CourseSmart. The platform was able to track each student's access patterns including
duration, frequency, time of access, and a proprietary metric called Engagement Index. Using these factors along with simple student demographic information and GPAs, the researchers performed blocked hierarchical linear regression to predict course grades, the subsequent model had an $R^2$ value of 0.254\textsuperscript{[6]}. These findings are significant in that they suggest both that LMS usage statistics have the potential to develop proxy measurements of student behaviors, and that student behaviors can be used as predictors of student performance.

These findings are further supported by research on predicting outcomes of online courses. Zacharis found that of the 29 different usage statistics recorded by the LMS he was examining, just 4 were needed to predict 52% of the variance in student grades\textsuperscript{[20]}. However, the variables used were only summaries of each student’s usage statistics throughout the course, thus they were not an accurate representation of the students’ week-to-week behaviors. Furthermore, it is important to acknowledge that the courses in that study were entirely online, not blended courses like those being examined in this study.

Outside of measuring student behaviors that contribute to performance, there are also easily accessible sources of quantitative data that can be used in prediction models. One source that is consistently used by students and professors alike is scores on assignments. Olama et al. examined graded assignments given within the first three weeks in a collection of four courses\textsuperscript{[19]}. These courses were selected such that they all had similar distributions of final grades, similar rubrics for grading, and similar time-lines for assignments. Each of the selected classes has three weekly graded assignments: a homework assignment, a quiz, and a discussion question. Using only these nine factors, both logistic regression and feed-forward neural networks were used to predict whether the students would pass or fail the course. Each week, the cumulative collected data was analyzed resulting in increasingly accurate predictions. The results of the two modeling methods were comparable. The logistic model's prediction accuracy rising from 80% to 87% to 91%, and the neural networks accuracy increasing from 81% to 85% to 89%\textsuperscript{[19]}. These findings demonstrate that the study habits and level of understanding becomes prevalent early on in a course, meaning grades on early assignments can be a strong predictor of final course grades. Unfortunately, this method of modeling student success fails to provide insights into students learning habits beyond the completion of graded assignments, and is only feasible
for regimented classes with frequent and consistent graded assignments.

Other research has ventured to predict students’ final course grade even before they set foot in the classroom by using previous scholastic achievement. In a 1992 study, Danko-McGhee and Duke modeled student success in an intermediate accounting class using their grade in an introductory accounting course and the number of semesters that had passed between being enrolled in the two classes as input variables. These two features were determined to be more statistically significant than the other variables included in the model: cumulative GPA and score on a diagnostic exam. Overall, the study’s linear model explained 41.6% of the variance in students’ grades for the intermediate accounting class [21]. While the study did not result in an exceptionally accurate model, it successfully shows the predictive power of student's past performance for their future performance indicating it as a potentially valuable factor in a more comprehensive model.

While variable selection is critical to predictive modeling, there are many different statistical techniques that can be employed to generate the actual model. Past research in predicting student performance has shown the usefulness of various methods, depending on the response variable being modeled and the intended use of the models. Fang & Lu showed the merit of a structured decision tree in predicting the letter grade of a course in a 2010 study [22]. Multiple linear regression is an often used method when predicting student grades on continuous scales, for it’s simplicity and interpretability [11, 13, 20]. However, most aligned with the goal of this research is logistic regression, with Macfayden and Dawson demonstrated as an effective means to model “risk of failure”, in a proof of concept for an “early warning system” [23].

This study will use the findings of past works and a basic understanding of modern learning theory to select variables to be collected and later used in modeling. Based on successful implementations of various modeling methods, Logistic Regression was selected for it’s ease of interpretation and ability to predict a binary response variable with a limited number of data points. Employing LMS usage data and individual assignment scores as predictors of course grade allows for continually updated predictions that does not require repeated intrusive data collection, with the added benefit of quantifiable increase of confidence in the predictions.

Amongst the numerous studies modeling course performance, there have been several
that attempt to do so across multiple classes and even subject matters. Recently, Gasevic et al. showed that predictors of success not only vary by course, but when universal models are made for multiple courses new predictors show significance [24]. These findings highlight the ease of overfitting models to the learning process. To help mitigate overfitting, and forced significance of variables, the number of variables included in the model was extremely limited and many that were included were known to have correlation with the final course grade.

3 Data and Limitations

This section will outline precisely how data was collected for this study, how that data was cleaned, and a brief discussion of the challenges associated with some of the data. Limitations of the metrics collected from each source are discussed in the subsections when relevant. Visualization of the various data sets is presented, along with interpretations of those visualizations. The data analyzed in this research came from two sections of a cross-listed engineering course at a large private institution in the North Eastern United States. The sections were taught by the same instructor however the syllabus and course structure did change between sections. The sections were taught in the Fall semesters of 2014 and 2015.

3.1 Learning Management System Data

The first data source examined in this study was the learning management system (LMS). Within this research, the LMS data collected primarily pertains to assignments within the class being analyzed. For each assignment, the number of submissions a student made (Revs) and the time elapsed between the first and subsequent submissions (RevTime) were interpreted as theoretical measures of how much a student revised their own work. In a similar method, the number of hours before a deadline that a student made their first submission for an assignment was treated as a measure of procrastination (Procrast). Also recorded was whether or not a student read any feedback posted on their assignments. This variable was measured as a percentage of the available feedback viewed at any given day (Feedback). Unfortunately, the 2014 section of the course did not distribute feedback digitally, and thus the measure was unavailable for that section.

LMS logs were used to collect metrics associated with time and frequency in which
students view course material. The metrics included the first date of access, the number of accesses, the last access, and the total duration of access for any given set of course notes, handouts, or other electronically distributed information. While potentially valuable, this information was stored cumulatively rather than periodically with time stamps. Consequently, there were no accessible measures of access patterns for, or up to, a specific week. Instead, only the summaries of the access patterns for the entire semester were available. For this reason, it was impossible to accurately derive the patterns of course access after the course was completed. Since this research aims to provide a method to identify students in need of feedback during a course, reflexive summary statistics of access patterns could not be used.

Table 1 shows the minimum, maximum, and median values for each of the LMS variables divided by course section. These summary statistics show the similarities and differences in these variables between sections. Note that the median and minimum values remain consistent across the sections while the other values change. This suggests that the typical student in each section is similar, but the outliers of each section are not consistent.

Table 1: Contained in this table is the summary statistics of data collected from the LMS for both the 2014 and 2015 sections of the course being studied.

<table>
<thead>
<tr>
<th></th>
<th>2014 Section</th>
<th></th>
<th>2015 Section</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Revs</td>
<td>RevTime</td>
<td>Procrast</td>
</tr>
<tr>
<td>Min.</td>
<td>1.00</td>
<td>0.00</td>
<td>-1.29</td>
</tr>
<tr>
<td>Max.</td>
<td>1.91</td>
<td>8.88</td>
<td>1.03</td>
</tr>
<tr>
<td>Mean</td>
<td>1.12</td>
<td>0.44</td>
<td>0.01</td>
</tr>
<tr>
<td>Median</td>
<td>1.00</td>
<td>0.00</td>
<td>0.01</td>
</tr>
<tr>
<td>Std.Dev.</td>
<td>0.18</td>
<td>1.54</td>
<td>0.31</td>
</tr>
</tbody>
</table>

Beyond the LMS usage data aggregated for this study; assignment grades were collected from the online learning environment. However, because the course assignment grades are available to students and professors by other means, this data was treated as a separate data source. Assignment grades were time-stamped so that models could accurately represent the data to the date they were generated. Despite the analyzed courses being the same subject matter and instructor, the number of assignments and their content varied. To help generalize the outcome of this study, individual assignment scores were not used as variables when modeling student success; rather, they were used to generate new variables that might better apply to new courses. This will be further discussed in Section 4.
### 3.2 Student Information System Data

In addition to the LMS data, data regarding the academic record of each student was collected from the university’s student information system (SIS). The course being analyzed for performance prediction required completion of two prerequisite classes for enrollment, “Introduction to Probability” and “Introduction to Statistics”. The final letter grade (A, B, C, D, or F) was collected, along with the academic term in which it was completed. Students that completed the prerequisite courses at another institution, before matriculation or otherwise, were marked with a grade of T. In conjunction with the grade each student earned in these prerequisite courses, the number of days from the completion of the prerequisite courses to the start of the course under study was measured. These variables (Prob.Gap and Stats.Gap) served as a measure of how “fresh” these fundamental topics were in the students’ minds. The student at status gap 0 was enrolled in the mandated prerequisite while taking the course under study. These variables are summarized in *Figure 1*.

*Figure 1*: Distribution of letter grades and time of enrollment in "Introduction to Statistics" and "Introduction to Probability" for students in both the 2014 and 2015 sections of the course under study.
As Figure 1 indicates, the 2014 section of the course had more uniform and consistent grades across the two prerequisite courses than students in the 2015 section. Figure 1 also indicates that most of the students in the 2014 section enrolled in the prerequisites at the same time, suggesting that they may have been enrolled in the courses with each one another. The histograms also show that the 2015 section earned lower letter grades in “Introduction to Statistics” than they did in “Introduction to Probability”. Most students in the 2015 section took “Introduction to Probability” at the same time, whereas their enrollment in “Introduction to Statistics” was much less consistent.

In addition to information regarding prerequisite courses, general academic statistics were collected. The variables recorded were: cumulative grade point average (CGPA), to represent general academic aptitude; enrolled credit hours during the term of the course being studied (CrsLoad), to capture the rigor of their other academic commitments; total completed credit hours up to the term of the studied course (CrdHrs), and the term when each student first matriculated into the university (FirstTerm), to quantify academic experience.

Table 2 shows the minimum, maximum, and median values for each of the variables from the SIS pertaining to general academic history. These summary statistics show the similarities and differences between these variables in different sections. Note that the mean and median values are similar between the sections, while the minimum, maximum, and standard deviations show greater differences. This indicates that the 2015 section had a larger variety of students, while still remaining similar to the 2014 section on average.

<table>
<thead>
<tr>
<th></th>
<th>2014 Section</th>
<th></th>
<th>2015 Section</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>CrsLoad</td>
<td>CrdHrs</td>
<td>CGPA</td>
<td>CrsLoad</td>
</tr>
<tr>
<td>Min.</td>
<td>6.00</td>
<td>17.00</td>
<td>2.45</td>
<td>6.00</td>
</tr>
<tr>
<td>Max.</td>
<td>20.00</td>
<td>147.00</td>
<td>3.92</td>
<td>18.00</td>
</tr>
<tr>
<td>Mean</td>
<td>15.00</td>
<td>116.00</td>
<td>3.39</td>
<td>15.00</td>
</tr>
<tr>
<td>Median</td>
<td>14.21</td>
<td>110.40</td>
<td>3.36</td>
<td>14.56</td>
</tr>
<tr>
<td>Std.Dev.</td>
<td>4.13</td>
<td>26.53</td>
<td>0.35</td>
<td>3.39</td>
</tr>
</tbody>
</table>

Figure 2 shows the term of matriculation for each student in both sections of the course being studied. There are very few students who matriculated in the Spring semesters, as expected, due to the academic year traditionally beginning in Fall. The majority of the students in the 2015 section matriculated one year after the majority of the
students from the 2014 section, indicating similar academic timelines between students in either section. Furthermore, the matriculation dates for students in the 2014 section follow a more predictable pattern than those for the 2015 section. However, after examining the raw data, CrdHrs was taken as a less biased indicator of academic experience than a chronographic measure, such as the number of months as a matriculated student, because of the irregularities in collegiate schedules, such as transfer students and graduate students.

In order to quantify the rigor of a student’s semester, the number of credit hours they were enrolled in during the course was recorded as a variable (CrsLoad). While credit hours may not fully quantify the difficulty of a course, alternative options had similar shortcomings and introduced new sources of bias. One alternative approach was to weight the credit hours of each course by the course number. This approach would consider a three credit 100-level course to be a lesser workload than a three credit 400-level course. However, to implement such a weighting system would require a (possibly arbitrary) evaluation of the relative difficulty of different course levels and implies that higher-level courses are always more demanding than lower-level courses, an assumption without empirical support.

3.3 Survey Data

The final source of data utilized in this study was an in-person survey distributed to students. This survey included 15 questions that attempted to quantify information descriptive to each student’s situation and study habits not available by existing digital means. The survey, in
its entirety, is available in the Appendix. The survey distribution was limited to only courses that took place during 2015, when this study commenced. There is no survey data available for the first section of the course under study, thus the survey data was not used in the final predictive modeling, but it’s analysis provided some valuable insights into student behaviors.

*Figure 3* shows the students ranking of their interest in the course, perceived relevance of the course, and previous knowledge of the subject matter as determined by their survey responses. These charts show that most students enrolled in the course with at least a basic understand of the material. It also shows that perceived relevance was has a fairly symmetric distribution with the mean falling between 3 and 4, suggesting that the students see some value in learning the material. Lastly, the histograms show the interest in the subject matter was centered on 4 with very few students showing little to no interest in the material. This suggests that most of the students would derive some internal motivation to learn the material.

The survey also asked students to self-report their allocation of time to both studying and extracurricular activities, including work and athletics. *Figure 4* shows each students allocation of time in hours, sorted from left to right by time dedicated to extracurricular activities. This plot suggests that there is no clear correlation between the time that a student spends studying and the time that they spend dedicated to extracurricular activities. Furthermore, no students spend more than ten hours a week studying for the course, and that extracurricular commitments range from 0 hours a week to the equivalence of a full-time job at 40 hours per week.
Questions about the students’ habits specific to the course under study were also asked in the survey. Students were asked to self-report their class attendance, completion of the suggested readings, preference to study with a group or individually, and whether they had access to the course materials prior to the class’s start. The results to these questions are illustrated in Figure 5. These charts show that nearly half of the students never did any of the suggested readings for the course and very few students did every reading. Students’ preferences for studying independently were evenly split with 34% preferring groups, 39% preferring studying alone, and the remainder having no preference. A significant portion of the class, 29%, had access to course materials before the course began.

Figure 4: Student time allocation to studying, as compared to extracurricular activities.

Figure 5: Student habits with regard to the analyzed course. From left to right the histograms show responses to questions about the frequency of class attendance, when suggested readings were completed, preferences to study in a group or independently, and if students had access to course materials before the course began.
Lastly, the survey attempted to measure self-advocacy in the course. To do this, students were asked how often they asked the teaching assistants and instructor for help. The responses are plotted in Figure 6. These plots show that most students did not go to teaching assistants for help, and those that did, did so rarely. Students were much more likely to ask for help from the instructor. The histogram shows that only seven students never got help from the professor and more than half the class went to the professor for help at least occasionally. This indicates students believe the professor was both willing and able to help them better perform in the class, giving hope for the effectiveness of an intervention.

![Figure 6: Summary of the frequency at which students request help from both the teaching assistant and professor in the course being studied.](image)

### 3.4 Data Cleaning and Challenges

Once data from all four of these sources was collected and organized into data tables, it was checked for completeness. Three students who did not have values for the majority of the variables were discarded, as they could not be modeled with the same dimensionality as the other students. Once complete cases were gathered, each of the numeric features was scaled to allow for comparison. When predefined bounds to the variables existed, such as exam grades (0-100), they were utilized. In all other cases, such as CourseLoad, the maximum and minimum observed values in the section were used as the bounds. Table 3 presents a list of all variables collected for
Outside of the challenges in interpreting the various metrics collected, this study was hindered by a policy change at the university that occurred within the span of the study. Between 2014 and 2015 the university changed from a letter grading scheme of A, B, C, D, and F to one that includes pluses and minuses, A, A-, B+, B, etc. To make the scheme consistent across all academic terms, pluses and minuses were removed from letter grades, whether they appeared for prerequisites or for the course under study itself. In rounding these letter grades, some information contained within the data was lost, however it was not possible to retroactively add pluses and minuses to already completed courses because numeric course grades were not kept on record. Along with the change in grade scheme, the university switched from a four-term schedule (Fall, Winter, Spring, Summer) to a more traditional three-term schedule (Fall, Spring, Summer). For consistency within this study, all credit hours were converted to reflect their weight in a three-term system, using the institution’s own method.
Table 3: This table contains all of the variables collected for this study, as well as their abbreviations and brief explanations.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Name</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>LMS</td>
<td>Revs</td>
<td>Average number of submissions per assignment.</td>
</tr>
<tr>
<td>LMS</td>
<td>RevTime</td>
<td>Average time between first and last submission for each assignment.</td>
</tr>
<tr>
<td>LMS</td>
<td>Procrast</td>
<td>Average time between each assignment deadline and first submission.</td>
</tr>
<tr>
<td>LMS</td>
<td>Feedback</td>
<td>Percentage of assignment feedback published online that was read by the student.</td>
</tr>
<tr>
<td>SIS</td>
<td>ProbabilityGrade</td>
<td>The letter grade the student earned in &quot;Introduction to Probability&quot;</td>
</tr>
<tr>
<td>SIS</td>
<td>StatisticsGrade</td>
<td>The letter grade the student earned in &quot;Introduction to Statistics&quot;</td>
</tr>
<tr>
<td>SIS</td>
<td>Prob.Gap</td>
<td>The number of days between the start of the course and the end of</td>
</tr>
<tr>
<td>SIS</td>
<td>Stats.Gap</td>
<td>&quot;Introduction to Probability&quot;</td>
</tr>
<tr>
<td>SIS</td>
<td>CGPA</td>
<td>The student’s cumulative grade point average before completing the course.</td>
</tr>
<tr>
<td>SIS</td>
<td>CrsLoad</td>
<td>The number of credit hours the student was enrolled in at the time of</td>
</tr>
<tr>
<td>SIS</td>
<td>CrdHrs</td>
<td>taking the course.</td>
</tr>
<tr>
<td>SIS</td>
<td>FirstTerm</td>
<td>The total number of credit hours completed by the student at the start of</td>
</tr>
<tr>
<td></td>
<td></td>
<td>the course.</td>
</tr>
<tr>
<td>SIS</td>
<td>Interests</td>
<td>The student’s self reported Interest in the course.</td>
</tr>
<tr>
<td>SIS</td>
<td>Relevance</td>
<td>The perceived relevance of the course to the student.</td>
</tr>
<tr>
<td>SIS</td>
<td>Knowledge</td>
<td>The student’s self reported prior knowledge of the course's subject matter.</td>
</tr>
<tr>
<td>SIS</td>
<td>ExtraCurr</td>
<td>The number of hours the student dedicates to extracurricular activities on</td>
</tr>
<tr>
<td>SIS</td>
<td>StudyHours</td>
<td>a weekly basis.</td>
</tr>
<tr>
<td>SIS</td>
<td>Attendance</td>
<td>The student’s self reported attendance to the course.</td>
</tr>
<tr>
<td>SIS</td>
<td>Reading</td>
<td>The student’s self reported frequency of reading suggested materials.</td>
</tr>
<tr>
<td>SIS</td>
<td>StudyPreference</td>
<td>The student’s preference, or lack thereof, to study individually or with a group.</td>
</tr>
<tr>
<td>SIS</td>
<td>PriorAccess</td>
<td>Whether the student had access to course materials before the class began.</td>
</tr>
<tr>
<td>SIS</td>
<td>ProfHelp</td>
<td>The student’s self reported frequency of seeking help from the instructor.</td>
</tr>
<tr>
<td>SIS</td>
<td>TAHelp</td>
<td>The student’s self reported frequency of seeking help from the teaching assistants.</td>
</tr>
</tbody>
</table>

4 Methodology

This section contains the methods and tools used to analyze the data described in the previous section. The first subsection details the explanations of how and why new variables were calculated for the student success models. The second subsection explains what modeling
techniques were employed to create predictive models discussed in the results section.

4.1 Variable Creation

In past research such as those presented in the literature review, the typical response variables for student performance modeling are numeric grade, 0-100; the categorical letter grade; A, B, C, D, or F; and the Pass/Fail binary outcome. While these responses are accurate to the empirical metrics of student success, such approaches rely on instructors determining which response values require intervention. To eliminate the need for instructors to choose performance thresholds for each student, this research engineered a response variable based on the students past academic performance.

The response variable developed was a binary measure indicating whether or not the student’s cumulative GPA (CGPA) would be negatively affected by the grade they earned in the course being modeled. To do this, the course grade, as a percentage, had to be converted to a 4-point scale. Typically, percentages are converted first to a letter grade, then to the corresponding number on the scale from 0 to 4. However, it is not uncommon for instructors to adjust a letter grade when converting it from a percentage. In other words, an instructor might give a student that scored an overall grade of 89/100 in the course an A during one semester and a different student who scored an 89/100 a B during a separate semester. Additionally, different instructors might have different definitions failing percentage. For courses where students typically earn higher grades the failing percentage might be 75%, and in courses where students typically earn lower grades the failing percentage might be 50%.

To generalize the transformation from a percentage to a four-point scale, a linear relationship was assumed starting with the failing percentage at 0 and ranging to 4 at 100%. The scaled grade was then multiplied by the credit value of the class and calculated into the student’s CGPA. If the new CGPA was lower than the original CGPA by 0.04 or more points, then the response (AffectedGPA) was 1. Other conditions impacting AffectedGPA are if the scaled grade was less than 1, resulting in a value of 1, regardless of CGPA; and a course grade being greater than 3, resulting in a value of 0. The population of students earning a converted grade greater than 3 is equivocal to the population of students that earned A’s. Not only can these students not earn a higher letter grade, but based on the results of the student survey they were already
Communicating with the instructor on a regular basis.

*Equation 1* shows how the course grade is converted to a four-point scale, depending on what percentage is determined to be a failing grade.

\[
\Psi = \frac{\text{CourseGrade} - \text{Failing Grade}}{100 - \text{Failing Grade}}
\]

*Equation 1*

The scaled course grade, \(\Psi\), is then weighted by the number of credit hours the course is worth and averaged with the students CGPA weighted by the total number of credits they have completed.

\[
\text{CGPA}_{\text{new}} = \frac{(\text{TotalCredits} \times \text{CGPA}_{\text{orig}}) + (\Psi \times \text{CourseCredit})}{\text{TotalCredits} + \text{CourseCredit}}
\]

*Equation 2*

The new CGPA was then compared to the original CGPA. If there was a decrease greater than the determined tolerance, or the scaled course grade was below 1, then \(\text{AffectedGPA}\) was 1. If the scaled course grade is greater than 3, then \(\text{AffectedGPA}\) was 0.

\[
\text{AffectedGPA} = ([\Psi < 1] \lor [(\text{CGPA}_{\text{orig}} - \text{CGPA}_{\text{new}}) > \text{Threshold}]) \land [\Psi < 3]
\]

*Equation 3*

While this method does programmatically create a response unique to each student, it still requires a manually set tolerance for an acceptable decrease in CGPA, which presents an opportunity for error by the instructors. However, rather than using an entirely arbitrary range, one could potentially be selected empirically based on scores in past sections. To do this, the differences between \(\text{CGPA}_{\text{orig}}\) and \(\text{CGPA}_{\text{new}}\) were plotted in ascending order. In *Figure 7*, the graphs show that a majority of CGPA decreases are quite small, however at a certain point the CGPA decreases start to become drastically larger. That point at which the slope of the plotted curves begins to increase dramatically was taken as the tolerance for a “significant decrease” in CGPA.
A sensitivity analysis was performed to determine how changes in the conversion from grades as a percentage to a letter grade would affect the changes in students’ CGPAs. Adjusting the lower limit for a passing grade from 50% up to 75% increased both the number of students who had their CGPA decreased and the amount by which their CGPA decreased. Knowing this, the range for what decrease in CGPA is acceptable could be changed when generating predictive models. The calculated CGPA decreases are plotted in Figure 7, showing the increased effect on students CGPA as the threshold for failure is increased.

As mentioned in Section 3, individual assignment grades were not used as variables when modeling student success. Rather, assignment grades were used to generate new variables to be utilized for modeling. This was done to lower the chances over fitting a model to a specific class or section of a class. The first variable created regarding student grades is a simple running average ($HWAvg$). This variable is easily calculated and represents a student’s general performance on assignments, rather than their understanding of specific material. The standard deviation of homework grades was also calculated as a potential variable for explaining student performance ($HWSD$). This feature would theoretically represent how consistently each student performed across different units of the course.

Taking these engineered variables further, in an attempt to make a more accurate representation of how a student is performing on homework more advanced algorithms were applied to the grade data. The existing grades were used in a moving average model that predicted what their homework average would be after 16 weeks, the total length of the course. The resulting value for each student was used as a variable when modeling student success ($ProjectedHW$). However, the moving average modeling method was unstable when less than
four homework grades were available for training the model to create the projected homework variable. This limited the variable’s existence to approximately halfway through the course. Using the same modeling technique, a second feature was extracted. This feature was developed with the aim of replacing \textit{HWSD}. Its calculation is simply the margin of error for the prediction generated by the moving average model (\textit{ProjectedError}). This variable was expected to be a measure of how unstable the student’s homework average was, thus indicating students that might see a sudden drop in homework average.

Another new variable was created based on the grades students earned in the prerequisite courses. With the sample size being small relative to the number of variables being considered for modeling, over fitting was a concern. To combat this, the variables \textit{ProbabilityGrade} and \textit{StatisticsGrade} were changed from categorical variables with five factors to binary variables indicating whether an A was earned in the class, reducing the dimensionality of the model by eliminating six possible dummy variables. Further manipulation was done to make a variable less specific to the course being studied. To do this, the sum of the two variables was calculated and divided by two. This new variable, A\textsubscript{inPrereqs}, is the portion of the prerequisite courses that the student earned an A in.

The last engineered variable was a basic summation of two other variables. This was done in order to lower the dimensionality of the models and prevent the use of two highly correlated variables in the model. Reducing the dimensionality also decreases the chances of over-fitting and improving predictive performance on new data. The variables combined were \textit{Prob.Gap} and \textit{Stats.Gap}. Due to the nature of these variables, a summation did little to lessen the information they contained. Separately, the variables were pseudo measures of how recently students had been exposed the subject matter of each prerequisite class. Combined, the new variable, Prereq.Gap, is a pseudo measure of how recently students have been exposed to prerequisite material in general. A listing of all the transformed variables is shown in Table 4, including how they are referenced in this text, their representation in the actual models, and a brief description.
Table 4: This table contains all of the variables developed in this study, as well as their abbreviations and brief explanations.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Name</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>eSIS1.1</td>
<td>Prob.Grade</td>
<td>A binary indicator of whether a student earned an &quot;A&quot; in &quot;Introduction to Probability&quot;</td>
</tr>
<tr>
<td>eSIS1.2</td>
<td>Stats.Grade</td>
<td>A binary indicator of whether a student earned an &quot;A&quot; in &quot;Introduction to Statistics&quot;</td>
</tr>
<tr>
<td>eSIS1.3</td>
<td>AinPrereqs</td>
<td>The portion of prerequisite classes that the student earned an &quot;A&quot; in.</td>
</tr>
<tr>
<td></td>
<td>Prereq.Gap</td>
<td>The sum of number of days between the end of the prerequisite courses and the start of the course being analyzed.</td>
</tr>
<tr>
<td>ePerf1</td>
<td>HWAvg</td>
<td>Student’s average grade on homework assignments.</td>
</tr>
<tr>
<td>ePerf2</td>
<td>HWSD</td>
<td>Standard deviation of student’s grades on homework assignments.</td>
</tr>
<tr>
<td>ePerf3</td>
<td>ProjectedHW</td>
<td>The student’s predicted final homework average using a simple moving average model.</td>
</tr>
<tr>
<td>ePerf4</td>
<td>ProjectedError</td>
<td>The margin of error from the moving average model for final homework average.</td>
</tr>
<tr>
<td>ePerf5</td>
<td>ExamAvg</td>
<td>Student’s average grade on exams.</td>
</tr>
<tr>
<td>Response</td>
<td>AffectedGPA</td>
<td>A binary indicator of whether the student's course grade lowered their CGPA by more than the determined threshold.</td>
</tr>
</tbody>
</table>

4.2 Analytical Techniques

This section details the modeling process used in this study. The data, collected and transformed, as described in Sections 3 and 4.1 respectively, was analyzed and used to train predictive models. Logistic regression, as described by McCullagh and Nelder in *Generalized Linear Models*, was selected as the method of modeling the binary response variable, AffectedGPA [25]. In order to identify which of the many variables were significant and strong predictors, stepwise logistic regression was employed, as outlined in *Modern Applied Statistics with S* [26]. In conjunction with the stepwise methodology further steps were taken to simplify and generalize the resulting models.

The data was broken into blocks based on the source it was collected from, the blocks were then further divided by the week in which the collected data represented. From here additional cumulative data sets were created, containing all of the individual weeks’ data. The cumulative data sets allow for a single model to be fit to the entirety of the course, rather than individual models for each week. With data from each source divided into slices of each week and a cumulative slice, it was then replicated to account for the variability of the response variable.
The response variable, $AffectedGPA$, is dependent on both the instructor’s standard for passing the course and the chosen tolerance for $CGPA$ decrease. To understand how those changes affect the response variable, a sensitivity analysis was performed on $AffectedGPA$. From the sensitivity analysis, five possible combinations of grading conditions were selected and applied to the replicated data slices. From here stepwise logistic regression was performed on each data slice independently.

With five models fit to each slice of data, there was notable variation between models suggesting that the models were not well generalized. To further generalize the models, the statistical significance of each variable in all of the models across each data source was examined. This resulted in eighty different models for each source of data, each with different coefficients and significant variables. From this list of eighty models, only the variables that appeared most frequently and consistently were kept, helping to ensure more generalized models.

The resulting fitted values from the logistic models, are the probabilities that the student will have a significant decrease in CGPA, indicated by a positive value of $AffectedGPA$. In order for that continuous response to be used to students as needing an “early alert” or not, thresholds for each model was set. For example, if the threshold is set at 0.5, then any predicted probability above 0.5 would receive a 1, and those at or below 0.5 would receive a 1. If these models were put into practice than the students who received a predicted change of 1 would be flagged for risk of under performance.

Type I versus Type II errors were examined in order to determine the most suitable probability threshold for classifying each student as having an $AffectedGPA$ or not based on the fitted logistic regression output. In this instance, the cost of unnecessarily alerting a student (Type I) is the additional work required from an instructor to issue the alert and follow-up with the student, as well as the additional time of the student. On the other hand, failing to notify a student in need of an alert (Type II) could lead to a permanent blemish on their academic record, having to retake the course, or potentially a delay in graduation. Considering the severity of these two errors, Type II errors were given twice the weight of Type I errors and they were combined into a single $Error Score$. 
Using the Error Score the classification threshold was calculated using three fold cross-validation. The data slices were randomly split into thirds, with two parts being used to train the model, and the remaining third being used to test the result. The fitted values for the test sets were then classified based on an arbitrary threshold, and the error score was calculated as described above. The threshold was incremented from zero to one, with the error score for each threshold being stored. This was then repeated 100 times, with different random samples being used for training and testing of the models. The threshold that resulted in the lowest average error score was selected as the optimal threshold for that given model.

5 Results

This section presents the results of the sensitivity analysis performed on the response variable (AffectedGPA), discusses the models that were trained for each source of data, and contains the results of the threshold optimization. Rather than presenting the models for every week, only the models generated for the weeks in which the institution requires “EarlyAlerts” are shown. The resulting models and thresholds are validated as to ensure logical and sensible results.

To ensure that the response variable is appropriate and rational, a sensitivity analysis was performed using the actual data. Taking the percentage grades earned by each student across all section of the course, the minimum passing grade, and tolerance for CGPA decrease were adjusted. For each combination of minimum passing grade and CGPA decrease, the number of students classified as having a positive value for AffectedGPA was recorded, Table 5 shows the results. There is a trend showing that when an instructor sets a higher standard for passing in a course, more students will experience decreases in CGPA and those decreases will be larger. As the tolerance for CGPA change was increased, the number of flagged students decreased exponentially, suggesting minor fluctuations in CGPA are common among students in the study.
Table 5: Number of students classified as having a significant decrease in CGPA based on the course’s minimum passing grade and the tolerance for CGPA decrease. Conditions that have been modeled in the remainder of this study are marked with an asterisk.

<table>
<thead>
<tr>
<th>Min. Passing Grade</th>
<th>Acceptable Decrease in CGPA</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.00</td>
</tr>
<tr>
<td>50%</td>
<td>41</td>
</tr>
<tr>
<td>55%</td>
<td>48</td>
</tr>
<tr>
<td>60%</td>
<td>53</td>
</tr>
<tr>
<td>65%</td>
<td>60</td>
</tr>
<tr>
<td>70%</td>
<td>66</td>
</tr>
<tr>
<td>75%</td>
<td>71</td>
</tr>
</tbody>
</table>

These groupings show that students in classes with higher standards for success should expect the potential for a larger negative impact on their CGPA. In the scope of this research, this table allows instructors to appropriately calibrate the range of an acceptable decrease in CGPA, based on what their standards of success are, and approximately how many students they feel capable of providing effective follow-up and feedback to. For example, if an instructor does not have the resources to provide meaningful additional support to 20 students, they may consider lowering the minimum passing grade, or anticipating the possibility of students seeing larger decreases in their CGPA.

While any one of these combinations of grading conditions is possible in practice, to move forward with the modeling process a minimum passing grade, and tolerance for CGPA decrease needed to be selected as to create a response variable for the training set. For the sake of completeness, five different combinations of minimum passing grade and acceptable range of CGPA decrease were used to create sets of training data. The combinations of minimum passing grade and tolerance for decrease in CGPA are indicated in Table 5 by asterisks. The models discussed in Subsections 5.1-5.4 were trained using the entirety of the available data as to most accurately generate coefficients. In Subsection 5.5 a more conservative sampling method was employed to explore the models’ potential value in predicting for new data sets.
5.1 Modeling Student Information System Data

Using data collected from the SIS, only the cumulative data was examined because the SIS data used in this study remained constant throughout the course. Initially using all the SIS and eSIS variables, the modeling process as described in Section 4.2 was followed. Including all of the SIS variables in stepwise regressions for each week of data, only the variables that most consistently showed statistical significance were kept for the models discussed in this section. This process of variable selection resulted in only two variables remaining, SIS$_3$ and SIS$_5$. The formulation of the model with the response variable based on the most lenient grading conditions is shown in Equation 4. This equation directly calculates the probability of a positive response variable.

\[
P(AffectedGPA) = \frac{100\%}{1 + e^{-(16.82 - 4.64 \cdot SIS_3 - 13.2 \cdot SIS_5)}}
\]

Equation 4

An alternative formulation of the same model is shown in Equation 5. This alternative form calculates a logistic odds ratio, or logit, rather than a probability. In doing so it makes for an easier interpretation of the variables and their coefficients. A lower logit indicates a smaller probability of a positive response for AffectedGPA and a larger logit indicates a larger probability. When the logit is equal to zero there is 50% chance of a positive response for AffectedGPA.

\[
\log \left( \frac{P(AffectedGPA)}{1 - P(AffectedGPA)} \right) = 16.82 - 4.64 SIS_3 - 13.2 SIS_5
\]

Equation 5

To interpret the model, consider Equation 5. The variables and their coefficients have a linear relationship with the logit. When SIS3 or SIS5 increases, the logit decreases, when SIS$_3$ or SIS$_5$ decreases, the logit value increases. Furthermore, it can be seen that a one unit change in SIS$_3$ has a smaller affect on the logit than an equivalent one unit change in SIS$_3$ based on the magnitude of their respective coefficients. This method of interpretation and the relationships between magnitude and direction of coefficients can be applied to all of the logistic models in this and the following sections.

The model in Equation 4, as well as the models for the four other grading conditions, are
shown in a tabular form in Table 6. Every iteration of the models have negative coefficients associated with both $SIS_3$ and $SIS_5$. This indicates that students with lower CGPAs and less completed credit hours are more likely to have a positive value for the response, AffectedGPA. This is inline with what was hypothesized, that student’s with lower CGPAs are more likely to earn a D, or a failing grade in the course. Likewise, students with fewer completed credit hours both have more easily decreased CGPAs based on the method by which it is calculated, and they have a less established academic performance, so even those with higher CGPAs may still be susceptible to an unexpectedly low grade.

Examining the AIC of the models for each set of grading conditions, it is clear that the instance in which the minimum passing grade was the highest and the tolerance for CGPA decrease was the lowest was the worst fit. This can be attributed to the large number of students being labeled with a true positive response variable. In the remaining instances the AICs are significantly lower, reaching its lowest value under the conditions in which the fewest students were labeled with a true positive response variable. Interestingly, the second model has a higher AIC than the fourth model despite having the same number of students labeled with true positives. This is likely attributable to more students receiving a D or F in the fourth model, where as the second model relied more on actual decrease in CGPA to flag students with a positive response variable.

*Table 6*: Logistic regression model using variables collected from the institution's student information system.

<table>
<thead>
<tr>
<th>Min. Passing Grade</th>
<th>CGPA Change</th>
<th>Tolerance</th>
<th>Week</th>
<th>Coefficients</th>
<th>AIC</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Intercept</td>
<td>$SIS_3$</td>
</tr>
<tr>
<td>50%</td>
<td>-0.09</td>
<td>All</td>
<td>16.82***</td>
<td>-4.64***</td>
<td>-13.20***</td>
</tr>
<tr>
<td>50%</td>
<td>-0.03</td>
<td>All</td>
<td>11.38***</td>
<td>-2.85***</td>
<td>-9.05***</td>
</tr>
<tr>
<td>60%</td>
<td>-0.06</td>
<td>All</td>
<td>28.01***</td>
<td>-7.26***</td>
<td>-17.21***</td>
</tr>
<tr>
<td>70%</td>
<td>-0.09</td>
<td>All</td>
<td>21.31***</td>
<td>-5.55***</td>
<td>-12.83***</td>
</tr>
<tr>
<td>70%</td>
<td>-0.03</td>
<td>All</td>
<td>9.80***</td>
<td>-2.49***</td>
<td>-4.16***</td>
</tr>
</tbody>
</table>

Statistical Significance codes: ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.1 ‘^’ 0.3

*Figure 7* shows how plotting the fitted values of the models in ascending order helps visualize their performance in isolating the true positive responses from the true negatives. The first and third models are the most accurate at identifying true positive response variables.
because they have the tightest groupings of true positive points. The marginal difference between the second and fourth models becomes more clear with the fourth model having fewer true positive observations with low fitted values. The first model, with the most lenient grading conditions, shows that none of the students with a fit probability of zero had a significant decrease in CGPA. However the nearly constant response variable suggests over fitting. The third model, with the median grading conditions, is similar although there is more variation in the response variables, suggesting that the over fitting is less dramatic. In the case of the second, a fourth, and fifth model there does not appear to be over fitting. The second model has a few true positive responses with low fitted probabilities, showing its inaccuracy. The fourth model successfully separates most of the true positive observations from the true negatives, with the exception of two observations that had relatively low fitted values. The fifth model has an indiscernible range of probabilities fit to true positive observations, indicating that it has little value for classification.

![Figure 7](image)

**Figure 7:** This figure shows the fitted values for models generated using only data from the student information system. Each plot represents a different model for a different set of grading conditions.

### 5.2 Modeling Learning Management System Data

The university’s learning management system data was used to create models. After performing stepwise regression across all of the slices of LMS data and comparing the results, only *Procrastination* was kept in the model because it consistently showed statistical significance. As seen shown in Table 7, all of the models resulted in negative coefficients for the *Procrastination*, which implies that students who submitted their assignments earlier, had a lower chance of having a negatively affected GPA. This is aligned with what was hypothesized, that students who do work early perform better than those who are rushed when completing their assignments. However, as indicated by the asterisks, or lack there of, *Procrastination* had fairly
high p-values in most models, with the exception of the cumulative models.

Table 7: The table below shows the models fit to week 7, week 12, and all the weeks of data collected from the learning management system. Asterisks indicate statistical significance of the terms, and the Akaike information criterion is included for each model.

<table>
<thead>
<tr>
<th>Min. Passing Grade</th>
<th>CGPA Change Tolerance</th>
<th>Week</th>
<th>Coefficients</th>
<th>AIC</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>Intercept</td>
<td>LMS&lt;sub&gt;3&lt;/sub&gt;</td>
</tr>
<tr>
<td>50%</td>
<td>-0.09</td>
<td>All</td>
<td>-1.64***</td>
<td>-5.75***</td>
</tr>
<tr>
<td>50%</td>
<td>-0.09</td>
<td>7</td>
<td>0.74</td>
<td>-13.93*</td>
</tr>
<tr>
<td>50%</td>
<td>-0.09</td>
<td>12</td>
<td>1.52</td>
<td>-12.86*</td>
</tr>
<tr>
<td>50%</td>
<td>-0.03</td>
<td>All</td>
<td>-1.59***</td>
<td>-1.68***</td>
</tr>
<tr>
<td>50%</td>
<td>-0.03</td>
<td>7</td>
<td>-0.84</td>
<td>-3.79^</td>
</tr>
<tr>
<td>50%</td>
<td>-0.03</td>
<td>12</td>
<td>-0.70</td>
<td>-3.33^</td>
</tr>
<tr>
<td>60%</td>
<td>-0.06</td>
<td>All</td>
<td>-1.75***</td>
<td>-1.66**</td>
</tr>
<tr>
<td>60%</td>
<td>-0.06</td>
<td>7</td>
<td>-1.46^</td>
<td>-2.34</td>
</tr>
<tr>
<td>60%</td>
<td>-0.06</td>
<td>12</td>
<td>-0.91</td>
<td>-3.19^</td>
</tr>
<tr>
<td>70%</td>
<td>-0.09</td>
<td>All</td>
<td>-1.67***</td>
<td>-1.41***</td>
</tr>
<tr>
<td>70%</td>
<td>-0.09</td>
<td>7</td>
<td>-1.42^</td>
<td>-2.03</td>
</tr>
<tr>
<td>70%</td>
<td>-0.09</td>
<td>12</td>
<td>-0.97</td>
<td>-2.68</td>
</tr>
<tr>
<td>70%</td>
<td>-0.03</td>
<td>All</td>
<td>-0.12</td>
<td>-1.53***</td>
</tr>
<tr>
<td>70%</td>
<td>-0.03</td>
<td>7</td>
<td>0.19</td>
<td>-2.34^</td>
</tr>
<tr>
<td>70%</td>
<td>-0.03</td>
<td>12</td>
<td>0.58</td>
<td>-2.77^</td>
</tr>
</tbody>
</table>

Notice that the coefficients for the model remain fairly consistent across all of the grading conditions except the two extreme conditions. When minimum passing grade is low and there is a high tolerance for CGPA decrease, then the coefficients are abnormally large. Plotting the fitted values generated by these models in ascending order shows the effect of the coefficients on predictions. The large coefficient results in a jump between two fitted probabilities, isolating only the most extreme case of procrastination and indicating the remaining observations as having a low probability of having a negatively affected CGPA.

On the other extreme, when the minimum passing grade was high and the tolerance for CGPA decrease was low, the models’ coefficients were abnormally small. This results in a very narrow range of fitted probabilities, with no notable distinction between true positive and true negative responses. This grading condition also resulted in the highest fitted values, as it is the condition under which the most students were labeled with a true positive response.
Figure 8 shows the fitted values produced by the models trained on the cumulative data collected for all of the weeks together. The upper row of graphs shows the fitted values when the models are applied to Week 7 observations and the lower row of graphs contains the fitted values produced by those same models applied to Week 12 data. The differences in the fitted values are due to the change in the measure of procrastination as weeks went on. With that said, the procrastination measure was an average so it became more stable in later weeks of the course.

Despite the coefficients of the individual week models plotted in Figure 8 and the cumulatively trained models plotted in Figure 9, the fitted values follow very similar trends. When comparing the individual week models to their cumulative model counterparts, the general trends are the same but the overall ranges of fitted probabilities are smaller in the cumulative models. This suggests that using a single model for all of the weeks of data could be used in place of one model for each week, without a significant loss of classification accuracy. While this is not immediately apparent when reading the AICs for the cumulative model and comparing to the AICs of the individual models, it is important to remember that the cumulative models are fit to fifteen times as many data points.
5.3 Modeling Grade Book Data

The same replication and modeling process was used on the grade book data source. This data changed as the course progressed so models were generated for both individual weeks and the cumulative data. From this data source, the variables that showed significance most consistently were \( HWAvg \) and \( ExamAvg \). This is unsurprising as they could be calculated from the moment that the first assignment was graded and the first exam was graded, respectively. Alternatively \( ProjectedHW \) primarily showed significance in later weeks of the course, likely due to its nature as a modeled response itself that increases in accuracy with additional data.

Displayed in Table 8, both \( HWAvg \) and \( ExamAvg \) have negative coefficients, meaning that students who earn higher grades are less likely to be labeled with a positive response variable. With the exception of the first grading conditions, where the minimum passing grade is 50% and the tolerance for CGPA decrease at 0.09, \( HWAvg \) has smaller coefficients than \( ExamAvg \). This implies that typically, exam grades better correlate with a student’s performance than their homework grades, however that relationship might switch given a different weighting schema for the course. The coefficients remained fairly consistent across the different grading conditions used to calculate \( AffectedGPA \), with exception to the two extreme conditions. These
differences are attributed to the number of students labeled with true positive response variables.

Table 8: The table below shows the models fit to Week 7, Week 12, and all the weeks of student performance data. Asterisks indicate statistical significance of the terms, and the Akaike information criterion is included for each model.

<table>
<thead>
<tr>
<th>Min. Passing Grade</th>
<th>CGPA Change Tolerance</th>
<th>Week</th>
<th>Coefficients</th>
<th>AIC</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>Intercept</td>
<td>ePerf₁</td>
</tr>
<tr>
<td>50%</td>
<td>-0.09</td>
<td>All</td>
<td>3.06</td>
<td>-5.25</td>
</tr>
<tr>
<td>50%</td>
<td>-0.09</td>
<td>7</td>
<td>3.29</td>
<td>-5.97</td>
</tr>
<tr>
<td>50%</td>
<td>-0.09</td>
<td>12</td>
<td>2.80</td>
<td>-4.74</td>
</tr>
<tr>
<td>50%</td>
<td>-0.03</td>
<td>All</td>
<td>10.05</td>
<td>-4.63</td>
</tr>
<tr>
<td>50%</td>
<td>-0.03</td>
<td>7</td>
<td>7.87</td>
<td>-4.64</td>
</tr>
<tr>
<td>50%</td>
<td>-0.03</td>
<td>12</td>
<td>13.05</td>
<td>-4.96</td>
</tr>
<tr>
<td>60%</td>
<td>-0.06</td>
<td>All</td>
<td>8.66</td>
<td>-5.18</td>
</tr>
<tr>
<td>60%</td>
<td>-0.06</td>
<td>7</td>
<td>8.10</td>
<td>-5.84</td>
</tr>
<tr>
<td>60%</td>
<td>-0.06</td>
<td>12</td>
<td>9.60</td>
<td>-5.01</td>
</tr>
<tr>
<td>70%</td>
<td>-0.09</td>
<td>All</td>
<td>10.20</td>
<td>-6.63</td>
</tr>
<tr>
<td>70%</td>
<td>-0.09</td>
<td>7</td>
<td>8.90</td>
<td>-7.21</td>
</tr>
<tr>
<td>70%</td>
<td>-0.09</td>
<td>12</td>
<td>12.83</td>
<td>-7.25</td>
</tr>
<tr>
<td>70%</td>
<td>-0.03</td>
<td>All</td>
<td>32.70</td>
<td>-14.44</td>
</tr>
<tr>
<td>70%</td>
<td>-0.03</td>
<td>7</td>
<td>25.17</td>
<td>-13.16</td>
</tr>
<tr>
<td>70%</td>
<td>-0.03</td>
<td>12</td>
<td>47.13</td>
<td>-16.55</td>
</tr>
</tbody>
</table>

Comparing the fitted values from the models generated using Week 7 data to the models generated using Week 12 data in Figure 10 shows a marginal improvement in classification. While this is beneficial, a more dramatic increase in the accuracy of the models was anticipated. This trend is illustrated by the graphs showing fitted values from the models trained based on a minimum passing grade of 70% and a tolerance for CGPA decrease of 0.09. The observations with true positive responses have higher fitted values in the Week 12 model than the Week 7 model. Simultaneously, the few observations with true negative response variables that had relatively high fitted values in the Week 7 model decreased in the Week 12 model, showing that classifications would become more accurate later in the course.
When comparing a single model trained across all of the weeks of data to the models trained for individual weeks, there is little difference in performance. The coefficients for the cumulative models split the difference between the coefficients for the models for Week 12 and Week 7. Plotting the cumulative models’ fitted values for Week 7 and 12, as in Figure 11, shows their performance in classifying students as having true positive response variables. While there is a certain degree of success, the models best identify students who earned Ds or failed the course, failing to classify those students who may be underperforming relative to their own academic record. Once again, the separation between the true positive responses and true negative responses only changes marginally between the fitted values from Week 7 to those for Week 12. However, being able to employ a single model across all weeks has the added benefit of simplicity, making results more easily interpreted and requiring less time for development.

Figure 10: These plots contain the fitted values for models generated using data collected on students’ in-class performance for individual weeks. The top row of graphs shows models fit to Week 7 data, whereas the second row shows models fit to Week 12 data; each column represents a different set of grading criteria.
Unfortunately, in both the individual week models and the cumulative models there is minimal separation between the observations that are true positives and those that are true negatives, with exception to the most extreme observations. This result shows that although there is a known correlation between assignment grades and final course grade, assignment grades alone are not enough to identify students who would benefit from feedback. Highlighting the need for leveraging the various sources that store student data.

5.4 Hybrid Modeling from Multiple Data Sources

To create a model that more completely represents student expectations and performance, the variables from all of the models discussed in Subsections 5.1-5.3 were employed in a single model. As anticipated, variables changed in significance and associated coefficients changed in magnitude. The procrastination variable, $LMS_3$, ceased bearing statistical significance and was removed from the model. Unsurprising, given its statistical significance was only marginal even when included in a model of its own. One variable changed in an unexpected way. $SIS_3$, experienced both a change in magnitude and direction, suggesting its relationship with the response variable was different from what previous analysis suggested.

Figure 11: These plots contain the fitted values for models generated using all of the data collected on students’ in course performance. The top row of graphs shows fitted values for Week 7 data, whereas the second row shows fitted values for Week 12 data; each column represents a different set of grading criteria.
Investigating the change in the coefficient’s direction, a strong correlation was found between $SIS_3$ and $ePerf_5$ (student exam average and student CGPA). In order to create a model that was not over-fit and easily interpretable one of these variables had to be selected for removal. In keeping with the objective of allowing for classification of students to be both accurate and early, $SIS_3$ was kept because of its immediate availability at the beginning of the course, whereas $ePerf_5$ remained undefined until the first exam was administered. This was done with negligible impact on the fitted probabilities generated by the models, further supporting the hypothesis that the use of both variables had caused over fitting during the modeling process.

Looking at the final models, as listed in Table 9, it is clear that the median three grading conditions resulted in relatively similar model coefficients. The first and fifth grading conditions however, have exceptionally large and exceptionally small coefficients respectively. The models for the first grading conditions, which has the fewest true positives, shows another oddity in that the AIC for the cumulative model is smaller than that of the Week 7 model, indicating a significant short coming in the Week 7 model. Under those same grading conditions, the Week 12 model has an AIC approaching zero, suggesting a near perfect model, which implies extreme over fitting.

Table 9: The table below shows the models fit to Week 7, Week 12, and all the weeks of data collected from each of the sources previously modeled. Asterisks indicate statistical significance of the terms, and the Akaike information criterion is included for each model.
Figure 12 shows plots of the fitted values generated by the Week 7 and 12 models in ascending order allows for an easy comparison of their accuracy. There is a noticeable improvement in the separation of variables with true positive responses, moving from Week 7 to Week 12. This is especially true when the tolerance for CGPA decrease is large and the standard for passing is low, but the previously discussed over fitting becomes even more apparent. The graph for the Week 12 model under these conditions shows perfect classification. Under the opposite grading conditions (low tolerance for CGPA decrease and high standard for passing), the improvement between weeks is less noticeable with significantly higher fitted probabilities being assigned to all of the observations.

Figure 12: These plots contain the fitted values for models generated using all of the data collected from the institution’s LMS. The top row of graphs shows fitted values for Week 7 data, whereas the second row shows fitted values for Week 12 data; each column represents a different set of grading criteria.

Examining the plots for models trained under the median grading conditions the improvement in model accuracy over time remains visible. Moving from Week 7 to Week 12, the fitted values for the observations with true positive responses increases, while the fitted values for observations with true negative responses decreases. This is observed in the graphs as a tighter grouping of true positives and a larger difference between that grouping and the closest true negative observation. This change is also visible in the models for the second and third grading conditions. While this change is significant and important to acknowledge, the Week 7
models are by no means inaccurate, supporting the goal of being able to identify students in need of feedback early on in the course.

To evaluate the cumulative models made using the combined data sources, under two plots were made for each set of grading conditions. The upper row of plots in Figure 13 shows the fitted values from Week 7 data being used in the cumulative models, and the lower row shows the fitted values from Week 12 data used in those same models. Once again, these plots show that the cumulative models are comparable in accuracy to the individual week models. In fact, under the first grading conditions, the cumulative model is more successful in identifying true positive responses in Week 7, and shows less overfitting in Week 12. These improvements, in conjunction with the benefit of having a single model for the duration of a course make the cumulative model much preferred to the individual models.

![Figure 13](image)

Figure 13: These plots contain the fitted values for models generated using data collected from all of the sources previously modeled. The top row of graphs shows models fit to Week 7 data, whereas the second row shows models fit to Week 12 data; each column represents a different set of grading criteria.

### 5.5 Optimizing Thresholds for Classification

If the purposed models were to be taken into practice then they must have established thresholds to convert the predicted probabilities into actual classifications. While these thresholds are not necessarily generalizable in themselves, they show the process and result of creating them. Thresholds were calculated for the hybrid model under each set of grading conditions.
conditions for individual weeks. To do this, threefold cross validation and optimization was performed, as described in Section 4.2, for Weeks 7 and 12 across all grading conditions. Plotting the resulting Type I and Type II error counts against the attempted thresholds shows how increases to the threshold affect classification errors. Figure 14 illustrates the relationship between threshold and the number of classification errors.

![Figure 14](image)

*Figure 14: These plots show the counts of Type I and Type II errors dependent upon the classification threshold for both week 7 and week 12 of each set of grading conditions.*

Under the most lenient grading conditions, with low minimum passing grade and large tolerance for CGPA decrease, there is little to no relationship between the threshold and the number of classification errors. This results in a very unstable optimal threshold that shifted from 0.18 to 0.98 from Week 7 to Week 12. The lack of relationship and instability are attributable to the small amount of true positive observations in that data set and the over fitting of the models. With so few positive responses to train the data on, the cross validation had wide margins of error and very few possible Type II errors.

The second grading conditions, with a low minimum passing grade and small tolerance for CGPA decrease, also saw a large change in the optimal threshold between Week 7 and Week 12. As opposed to being attributable to the model’s instability, in this case the shift in the optimal threshold is likely due to an observable improvement in the model over time. As shown in Figure 13, the clusters of true positive responses and true negative responses become more tightly packed so the threshold can be decreased without creating additional misclassification.
In the remaining set of models, under the other three grading conditions, the optimal threshold remains constant. This is because there is little to no change in the performance of the models between Weeks 7 and 12. This gives some guidance for future application of this methodology, in that if the models are not observed to be changing over time, then there is no need to recalculate the optimal threshold for classification.

6 Conclusions

This section contains the interpretation of the findings from this study and explains their implications. Additionally, the following paragraphs lay out possible next steps in continuing research in this field.

Of the results from this research, perhaps the most important is the ability to utilize noninvasive data to create models that helped predict student performance early on in a course. Additionally, the data that proved the strongest predictors of student performance were past course performance and current assignment grades. While other variables might have more significance in different courses, the variables outlined in this study can serve as a baseline or jumping off point if these models were to be replicated. Inspecting the variables that showed significance in the models from this study highlights the importance of a data collection from diverse courses. It is insufficient to only view past performance, current performance, or academic behavior independently when identifying students in need of additional feedback and support.

In conjunction with the need to collect data from multiple sources, the need for a response variable more advanced than a simple letter or numeric grade is apparent. Through distributed surveys a strong correlation between course grade and CGPA was discovered. In essence, the findings showed that “A students” get As and “C students” get Cs. While this is by no means groundbreaking, it does support the need for Early Alert criteria to stretch beyond course grade. If a model were to only predict course grade and flag students on track to fail, underperforming “A students” would rarely, if ever, be identified as in need of intervention. That being stated, it is also relevant to keep in mind that students who performed well in the courses considered in this study were already in regular contact with the professor. In other words, the strongest students were less in need of formal intervention like Early Alerts, not because of their grades, but because of their self-advocacy.
Furthermore, the sensitivity analysis performed on the response variable shows how instructors’ chosen grading schema can impact students. The observed increase in underperforming students was expected when the minimum passing grade rose. However, the inverse relationship between small tolerance for CGPA decrease and the number of students flagged was not anticipated to match it so directly. The sensitivity analysis shows that for instructors to provide a sufficient, but not unmanageable, amount of intervention they must balance their grading standards with their expectations for the course grades students will earn. Additionally, instructors with higher standards should be prepared and willing to provide more support to their students.

Looking beyond the relationships that were identified by this research, additional interesting findings were discovering that relationships expected to exist, actually did not. One such finding that was hinted to earlier in this section is the predictive power of the prerequisite course grades. Using only prerequisite grades to predict course grade is fairly accurate and a relationship between the two is seemingly obvious. However, once CGPA is included as well, the significance of prerequisite grades drops. After investigation, there is a strong co-linearity between CGPA and prerequisite grades, and CGPA seems to be the true predictor of success. This suggests that students in more advanced classes have established academic performance, or are enrolled in courses that reflect their subject matter expertise.

A variable that did not meet expectations in predicted performance is Procrastination. Recall that procrastination did have some success as a predictor of student success when used on its own in a model. Once other variables, such as assignment scores and academic history were included, procrastination no longer contributed to making the predictions more accurate. While this is somewhat disappointing, it is supported by recent learning theory, which claims that high performing students have similar behaviors, not that those behaviors make students high performers [27]. Further analysis into these findings could generate more empirical support for the arguments against behaviorism as a learning theory.

In future research, it could be beneficial to model individual student performance across their collegiate careers in order to better tailor feedback even better. This would allow for irregularities in individual behaviors to be identified, rather than deviations from the groups behaviors. Additionally, this study was limited by a small sample size and institutional changes that affected data collection. A larger study, over a more consistent academic period could have
more robust results than those presented in this study. With data from additional courses, over a longer period of time some of the questions that arose in this research might be answered.

One such question came about when examining the timing of prerequisite courses. When using the term in which a student completed a prerequisite course there were individual terms that correlated strongly with underperformance. However, they did not appear to be in any sort of pattern suggesting that it was a difference in something other than the timing that caused the correlation. Further investigating this sort of finding could be beneficial to a university in identifying weak instructors or syllabi, and potentially make other instructors aware of groups of underprepared students.

In a similar vein but on a smaller scale, looking at the significance of specific assignment grades within a course showed that some assignments were better at predicting student success than others. These findings may suggest that the topics of certain assignments are more foundational or more challenging than others, information likely helpful to instructors. However, using these scores as predictors of performance is surely not generalizable across different courses, and thus out of scope for this research.

Overall, the results of this research show that it is feasible to automate a large portion of the performance feedback process for professors. While the variables that were found to be strong predictors are not surprising, they are also not complex making the models easily interpreted. Implementing such a system could have a direct impact on students and professors, in an area already acknowledge by institutions as relevant. Additionally, utilizing such a system would provide a consistent baseline for measuring the effectiveness of early interventions with students, something fundamental to future research.

8 Acknowledgments

Foremost, I would like to thank my thesis advisor, Dr. Rachel Silvestrini. I would not have attempted this endeavor without her encouragement and could not have completed it without her endless patience and support. She consistently pushed me to better my research, writing, and self; for that I am humbly gracious.

I would also like to acknowledge all the professors and staff in the Industrial and Systems Engineering Department at RIT. Their dedication and commitment made my education both
challenging and truly rewarding. I owe a special thanks to Mark for always being there to help me take my mind off of classes and my thesis. I truly enjoyed our time together.

Lastly, I would like to thank my parents, brother, and girlfriend. Their love and unconditional support throughout my years of study helped make this possible. Without their phone calls, visits, and hours of editing I would not be here. Thank you.

9 References


10 Appendix

How would you rate your interest in the topics covered in this course?

<table>
<thead>
<tr>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Not Interested</td>
<td>Very Interested</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

How would you rate the relevance of the topics covered in this course to your desired profession?

<table>
<thead>
<tr>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Not Relevant</td>
<td>Very Relevant</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

How would you rate your prior knowledge of the subject matter covered in this course?

<table>
<thead>
<tr>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Very Little Prior Knowledge</td>
<td>Extensive Prior Knowledge</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

How often do you attend this class?

a. Always  
b. Regularly  
c. Rarely  
d. Only exams

Where in the classroom do you normally sit? Draw an 'X' in the box below that corresponds to the location in the classroom.

[Diagram of classroom with 'X' in the front]

On average, how much time per week do you spend studying for this course, outside of class?

_______ hours
In regards to this class, do you prefer to study with a group, or by yourself?
   a. I prefer to study in a group.
   b. I prefer to study by myself.
   c. I have no strong preference between the two.

Which one of the following statements best describes your behavior with regards to the readings assigned in this class?
   a. I do all the readings
   b. I read chapters that I’m struggling with.
   c. I read to supplement lectures I’ve missed
   d. I do not do any of the readings.

How many hours a week do you spend doing extracurricular activities (athletics, work, clubs)?
   ________ hours

How often do you get help from one of the TA's in this class?
   a. Never
   b. Rarely (≤3 times)
   c. Occasionally (~5 times)
   d. Regularly (every other week)
   e. Frequently (every week)

How often do you get help from the professor in this class?
   a. Never
   b. Rarely (≤3 times)
   c. Occasionally (~5 times)
   d. Regularly (every other week)
   e. Frequently (every week)

Did you have access to course materials prior to the start of this class?
   Yes / No

What letter grade did you receive in Introduction to Probability (ISEE 251)?

What letter grade did you receive in Introduction to Statistics (ISEE 252)?