A standards-based grading model to predict students' success in a first-year engineering course

Farshid Marbouti
Purdue University

Follow this and additional works at: http://docs.lib.purdue.edu/open_access_dissertations

Part of the Educational Assessment, Evaluation, and Research Commons, Engineering Commons, and the Higher Education Commons

Recommended Citation
http://docs.lib.purdue.edu/open_access_dissertations/675

This document has been made available through Purdue e-Pubs, a service of the Purdue University Libraries. Please contact epubs@purdue.edu for additional information.
This is to certify that the thesis/dissertation prepared

By Farshid Marbouti

Entitled
A Standards-Based Grading Model to Predict Students’ Success in a First-Year Engineering Course

For the degree of Doctor of Philosophy

Is approved by the final examining committee:

Heidi Diefes-Dux
Co-chair

Krishna Madhavan
Co-chair

Matthew Ohland

Joyce Main

To the best of my knowledge and as understood by the student in the Thesis/Dissertation Agreement, Publication Delay, and Certification Disclaimer (Graduate School Form 32), this thesis/dissertation adheres to the provisions of Purdue University’s “Policy of Integrity in Research” and the use of copyright material.

Approved by Major Professor(s): Heidi Diefes-Dux/Krishna Madhavan

Approved by: Ruth Streveler
Head of the Departmental Graduate Program

Date 1/19/2016
A STANDARDS-BASED GRADING MODEL TO PREDICT STUDENTS’ SUCCESS
IN A FIRST-YEAR ENGINEERING COURSE

A Dissertation
Submitted to the Faculty
of
Purdue University
by
Farshid Marbouti

In Partial Fulfillment of the
Requirements for the Degree
of
Doctor of Philosophy

May 2016
Purdue University
West Lafayette, Indiana
For my wife, my mother, and in the memory of my father.
ACKNOWLEDGEMENTS

I would never have been able to finish my dissertation without the guidance of my committee members, and support from my wife and family.

I would like to express my deepest gratitude to my advisors, Professor Heidi Diefes-Dux and Dr. Krishna Madhavan, for their excellent guidance, caring, patience, and providing me with an excellent atmosphere for doing research. Their guidance and feedback helped me to organize my thoughts and clarify my writing.

I would also like to thank my committee members, Dr. Joyce Main and Dr. Matthew Ohland, who supported me with their thoughtful feedback during this process.

I want to thank all of my family for their support throughout this process. My mother has always been a great support and encouraging me with her best wishes.

Finally, I would like to thank my wife. She was always there cheering me up and stood by me through the good times and bad. I couldn't imagine making this journey without her.
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>LIST OF TABLES</td>
<td>viii</td>
</tr>
<tr>
<td>LIST OF FIGURES</td>
<td>ix</td>
</tr>
<tr>
<td>ABSTRACT</td>
<td>xi</td>
</tr>
<tr>
<td>CHAPTER 1 - INTRODUCTION</td>
<td>1</td>
</tr>
<tr>
<td>1.1 Students’ success and retention rates</td>
<td>1</td>
</tr>
<tr>
<td>1.2 What is the solution?</td>
<td>2</td>
</tr>
<tr>
<td>1.3 Research purpose</td>
<td>4</td>
</tr>
<tr>
<td>1.4 Research questions</td>
<td>5</td>
</tr>
<tr>
<td>CHAPTER 2 - LITERATURE REVIEW</td>
<td>7</td>
</tr>
<tr>
<td>2.1 Grading systems: summative score-based vs. standards-based</td>
<td>7</td>
</tr>
<tr>
<td>2.2 Learning analytics and predictive modeling</td>
<td>11</td>
</tr>
<tr>
<td>2.3 Predict students’ success in a course</td>
<td>12</td>
</tr>
<tr>
<td>2.4 Generic early warning systems</td>
<td>13</td>
</tr>
<tr>
<td>2.4.1 Examples of existing early warning systems</td>
<td>14</td>
</tr>
<tr>
<td>2.4.2 Shortcomings of early warning systems</td>
<td>16</td>
</tr>
<tr>
<td>2.5 Course specific prediction systems</td>
<td>18</td>
</tr>
<tr>
<td>2.6 Significance of this research</td>
<td>23</td>
</tr>
</tbody>
</table>
CHAPTER 3 – METHODS

3.1 Data source and settings

3.2 Data

3.3 What is success in the course?

3.4 Prediction modeling methods

3.4.1 Logistic Regression

3.4.2 Support Vector Machine

3.4.3 Decision or Classification Tree

3.4.4 Multi-Layer Perceptron

3.4.5 Naive Bayes Classifier

3.4.6 K-Nearest Neighbor

3.4.7 Ensemble method

3.5 Train, verify, and test the models

3.6 Feature Selection

3.6.1 Correlations

3.6.2 Gini gain

3.6.3 Sum of Squared Errors (SSE) and explained variance ($R^2$)

3.6.4 Optimal number of variables

3.7 Model robustness

3.8 Workflow of model development
CHAPTER 4 - MODEL DEVELOPMENT........................................................................47

4.1 Training and verifying the models....................................................................47
  4.1.1 Misidentifications and errors.................................................................48
  4.1.2 Best models for predicting students who passed.................................49
  4.1.3 Best models for predicting students who failed.................................50

4.2 Creating an Ensemble model..........................................................................51

4.3 Predictive power of different assessments.....................................................53

4.4 Feature selection............................................................................................55

4.5 Model robustness...........................................................................................58

4.6 Testing the models performance.....................................................................59

CHAPTER 5 - LEARNING OBJECTIVES.................................................................65

5.1 Learning objectives used in the models.........................................................65

5.2 Learning objectives after week 5.....................................................................70

5.3 Top predictive learning objectives..................................................................74

CHAPTER 6 - DISCUSSION...................................................................................76

6.1 R.Q. 1 – Best prediction modeling method..................................................76

6.2 R.Q. 2 – Selection of student performance data..........................................84

6.3 R.Q. 3 – Achievement of learning objectives................................................87

6.4 Recommendations for creating course specific models..............................91

6.5 Limitations.......................................................................................................95

6.6 Future research...............................................................................................97
6.7 Conclusion .................................................................................................................. 99

LIST OF REFERENCES ................................................................................................... 101

LIST OF REFERENCES ................................................................................................... 101

VITA ........................................................................................................................................ 109
# List of Tables

<table>
<thead>
<tr>
<th>Table</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1</td>
<td>STEM course specific prediction studies</td>
<td>20</td>
</tr>
<tr>
<td>2.2</td>
<td>Accuracy of the predictions using different prediction methods and variables</td>
<td>22</td>
</tr>
<tr>
<td>3.1</td>
<td>Course syllabus and assessments for the Spring 2013 and 2014 FYE course</td>
<td>28</td>
</tr>
<tr>
<td>3.2</td>
<td>Number of missing grades. Missing grades replaced by zero</td>
<td>29</td>
</tr>
<tr>
<td>3.3</td>
<td>Train, verification, and test datasets</td>
<td>36</td>
</tr>
<tr>
<td>4.1</td>
<td>Verification results and accuracy of prediction models</td>
<td>48</td>
</tr>
<tr>
<td>4.2</td>
<td>Test results and accuracy of the Ensemble model</td>
<td>53</td>
</tr>
<tr>
<td>4.3</td>
<td>$F_{1.5}$ scores of the models created with different assessment grades</td>
<td>54</td>
</tr>
<tr>
<td>4.4</td>
<td>Accuracy of predicting success at midterm exam</td>
<td>55</td>
</tr>
<tr>
<td>4.5</td>
<td>Test results of top two models using two variables selected by the correlation method</td>
<td>63</td>
</tr>
<tr>
<td>5.1</td>
<td>Description of homework learning objectives for weeks 1 to 5 that were used in the prediction models</td>
<td>66</td>
</tr>
<tr>
<td>5.2</td>
<td>Description of homework learning objectives after week 5</td>
<td>71</td>
</tr>
<tr>
<td>5.3</td>
<td>Top five learning objectives selected by feature selection methods</td>
<td>74</td>
</tr>
<tr>
<td>6.1</td>
<td>Guidelines for using prediction modeling methods</td>
<td>93</td>
</tr>
</tbody>
</table>
## LIST OF FIGURES

<table>
<thead>
<tr>
<th>Figure</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.1 ENGR 132 concept map of learning objectives</td>
<td>27</td>
</tr>
<tr>
<td>3.2 Distribution of letter grades for FYE course in the Spring 2013 and 2014 semesters</td>
<td>31</td>
</tr>
<tr>
<td>3.2 Correlation between predicting and predicted variables</td>
<td>39</td>
</tr>
<tr>
<td>3.3 Gini gain of the predicted variables</td>
<td>41</td>
</tr>
<tr>
<td>3.4 Explained variance for predicted variables</td>
<td>43</td>
</tr>
<tr>
<td>3.5 Overall workflow of model development</td>
<td>46</td>
</tr>
<tr>
<td>4.1 Number of misidentifications based on the letter grades</td>
<td>49</td>
</tr>
<tr>
<td>4.2 Number of times a student has been misidentified by the six models</td>
<td>52</td>
</tr>
<tr>
<td>4.3 Number of Ensemble model misidentifications</td>
<td>53</td>
</tr>
<tr>
<td>4.4 Accuracy of the models using different numbers of variables</td>
<td>56</td>
</tr>
<tr>
<td>4.5 Accuracy of the models using different number of variables</td>
<td>57</td>
</tr>
<tr>
<td>4.6 Accuracy of the models using different number of variables</td>
<td>58</td>
</tr>
<tr>
<td>4.7 Accuracy of the models based on different training size</td>
<td>59</td>
</tr>
<tr>
<td>4.8 Accuracy of the models for future data using different numbers of variables</td>
<td>61</td>
</tr>
<tr>
<td>4.9 Accuracy of the models for future data using different numbers of variables</td>
<td>61</td>
</tr>
<tr>
<td>4.10 Accuracy of the models for future data using different numbers of variables</td>
<td>62</td>
</tr>
<tr>
<td>Figure</td>
<td>Page</td>
</tr>
<tr>
<td>--------</td>
<td>------</td>
</tr>
<tr>
<td>4.11 Number of misidentifications for the top two models</td>
<td>64</td>
</tr>
<tr>
<td>5.1 Distribution and average learning objective scores for weeks 1-5 for students who passed and failed</td>
<td>67</td>
</tr>
<tr>
<td>5.2 Correlation among course assessments that were used in the prediction models</td>
<td>69</td>
</tr>
<tr>
<td>5.3 Correlations among homework learning objectives</td>
<td>73</td>
</tr>
</tbody>
</table>
ABSTRACT

Marbouti, Farshid PhD, Purdue University, May 2016. A Standards-Based Grading Model to Predict Students' Success in a First-Year Engineering Course. Heidi Diefes-Dux & Krishna Madhavan.

Using predictive modeling methods, it is possible to identify at-risk students early in the semester and inform both the instructors and the students. While some universities have started to use standards-based grading, which has educational advantages over common score-based grading, at-risk prediction models have not been adapted to reap the benefits of standards-based grading. In this study, seven prediction models were compared to identify at-risk students in a course that used standards-based grading. When identifying at-risk students, it is important to minimize false negative (i.e., type II) errors while not increasing false positive (i.e., type I) errors significantly. To increase the generalizability of the models and accuracy of the predictions, feature selection methods were used to reduce the number of variables used in each model. The Naive Bayes Classifier and an Ensemble model using a combination of models (i.e., Support Vector Machine, K-Nearest Neighbors, and Naive Bayes Classifier) had the best results among the seven tested models. This study identified possible threshold concepts and learning objectives that are important to students' success in the course, and learning objectives that are not correlated with student success in the course.
CHAPTER 1 - INTRODUCTION

"The first thing the (professor) told us was, 'You should expect to see this class dwindle down as the semester goes on.' It was the first thing they told us" (Boundaoui, 2011). This is the first thing Amaneh remembers from her first class in college. There are a lot of students like Amaneh who believe that the first year of undergraduate Science, Technology, Engineering, and Mathematics (STEM) programs are intentionally made difficult by the instructors to see who can make it and who cannot. Lesley, a freshman student says: “Chemistry 1 is also a weed-out class, meaning it’s extra difficult and meant to determine who is cut out for the pre-med life. I failed first semester” (Newlon, 2013). As Lesley mentions, it seems in most colleges the STEM programs, engineering included, especially in the first year, promote failure instead of success. It is then not surprising that STEM students migrate to non-STEM programs or quit college by the end of the first year.

1.1 Students’ success and retention rates

Despite numerous efforts to improve students’ retention and success in higher education institutions over the past 30 years, retention rates remain similarly low (Gardner & Koch, 2012). According to the National Collegiate Retention and Persistence to Degree Rates report in 2012, the first to second year retention
rate was 66.5% on average (ACT, 2012). Low retention rates reveal a serious problem in higher education. Almost one third of students leave college after experiencing their first year. The attrition continues through the next years of school - only 45% of students who enter college graduate after 5 years (ACT, 2012).

When it comes to low graduation rates, engineering students are not an exception. In a research study, Ohland and colleagues (2008) used Multiple-Institution Database for Investigating Engineering Longitudinal Development (MIDFIELD) and the results of the 2007 National Survey of Student Engagement (NSSE) to look at engineering graduation rates. According to this research, while engineering students are similar to other majors in terms of persistence in major, the rate of students changing their major to engineering is much lower (7%) than other majors (30%) (Ohland et al., 2008). In other words, engineering majors lose students similar to other majors, but do not gain as many students as others.

Based on the President’s Council of Advisors on Science and Technology report in February 2012, fewer than 40% of college students entering a STEM program complete a STEM degree (Gates & Mirkin, 2012). At the same time, there is a need for one million more STEM professionals over the next decade (Gates & Mirkin, 2012). An increase of 10% in the graduation rate in STEM degrees would generate three-quarters of the one million needed STEM professionals.

1.2 What is the solution?

Academic success is the most important factor in students’ retention and the best predictor of students’ persistence (DesJardins, Ahlburg, & McCall, 1999;
Pascarella & Terenzini, 2005). Risk of attrition decreases with an increase in academic achievement (Murtaugh, Burns, & Schuster, 1999). Thus, one way to increase retention is to increase academic success.

The problem of low success rate in higher education is two fold. While many efforts have focused on what to do to increase students' success, some scholars are questioning the validity of summative score-based grading, which is common in STEM courses, for measuring students' achievements (Carberry, Siniawski, & Dionisio, 2012; Marzano, 2010; Sadler, 2005).

Moving from high school to college is considered a big change in students' educational life and a lot of effort has focused on improving students' success in their first year of college. Adding first-year seminar courses (e.g., Miller, Janz, & Chen, 2007), orientation classes (e.g., Starke, Harth, & Sirianni, 2001), and summer programs (e.g., Vinson, 2008) are some of the efforts to help first-year students succeed. Creating living/learning communities (e.g., Purdie & Rosser, 2011), assistance programs such as those for socially/economically disadvantage students (e.g., Braunstein, Lesser, & Pescatrice, 2008), and remedial English programs (e.g., Leake & Lesik, 2007) are examples of programs that all college students may benefit from. While the creators of these programs are optimistic, the research results on whether or not these programs improve students' educational success are mixed. In other words, some studies show that these programs are beneficial for students while some conclude there is no difference in the academic success of students who participated in these programs and those who did not.
Invalid and unreliable grading systems may discourage students in a course, which consequently contribute to low success rate. Standards-based grading can be used as an alternative to the summative score-based grading. In standards-based grading, assessments measure the quality of students' work (i.e., students' achievement) on specific learning objectives regardless of other students' performance (Sadler, 2005). Standards-based grading has several educational benefits for students. Most importantly, it provides students with clear and meaningful feedback (i.e., formative assessment) related to achievement of the course learning objectives and helps students identify their weaknesses in the course (Atwood & Siniawski, 2014).

An alternative solution to help students be more successful in their courses and consequently in their program is to utilize the existing data in higher education institutes, collected by instructors during courses, to increase students' success. Instead of merely relying only on instructors' experience or anecdotal evidence, another way to learn how to enhance students' success is to analyze students' performance data (Huang & Fang, 2013; White, 2012) using learning analytics methods. The results of such analyses provide information for both instructors and students that can help instructors understand what leads to a student passing or failing a course and provide better ways to promote academic success.

1.3 Research purpose

It is crucial to predict students' success during the semester by identifying at-risk students early in the semester. In my dissertation research, I found the
relationships among achievement of the various course learning objectives and students' grades on the course assessments in a first-year engineering course. This highlighted which learning objectives and assessments are related with success in the course. These relations provided information that was used to create and optimize performance-based prediction models for the course. I used only academic factors including learning objectives achievement scores and assessment grades, which are available to the course instructor during the semester, and built prediction models to identify at-risk students in the course. The models were used as a proof of concept to showcase and move toward course-specific learning objective-based prediction models rather than the existing score-based early warning systems.

1.4 Research questions

1. Of six different predictive modeling methods, as well as a seventh hybrid or Ensemble method (consisting of three of the most successful individual methods, see section 3.4 for more details), which is the most successful at identifying at-risk students, based on specified in-semester student performance data? Why is this method the most successful? Why are the other methods less successful?

2. To what extent can the models created by predictive methods for identifying at-risk students in a course be improved through the selection of in-semester student performance data (e.g., quiz, homework learning objectives, midterm exam)? What does the selection reveal?
3. What are the relationships, if any, between students’ success and achievement of different learning objectives in a course? What are the implications for the resulting prediction models and what are the pedagogical implications?
CHAPTER 2 - LITERATURE REVIEW

In this chapter, I describe standards-based grading and its educational advantages, and then briefly review learning analytics and predictive modeling. Then I move into reviewing early warning systems and their benefits and shortcomings. This leads to a review of course specific prediction models for STEM courses. At the end of this chapter, I discuss the significance of this research study.

2.1 Grading systems: summative score-based vs. standards-based

In traditional grading systems, students’ success is measured based on how many points have been obtained during the semester. However, a point is an arbitrary unit of measurement, not a well-defined constant unit. The value of a point changes from one instructor to another, and for an instructor from one semester to another (Post, 2014). Thus, it is not surprising that the score a student receives on an assignment is less dependent on a student’s achievements than on who scores the assignment and how it is being scored (Marzano, 2010). The score-based system “gives the illusion of precision” (Post, 2014, p. 2), but the grades may not be connected to competencies in the course learning objectives (Sadler, 2005).

Defining course learning objectives at the beginning of the semester is part of an instructional design referred to as backward curricular design (Wiggins &
McTighe, 1998). This means instead of beginning with textbooks and favorite lessons to design a course, one starts with the desired outcomes (i.e., what we expect the students to learn or be able to do at the end of the semester) and work backward to design the curriculum and assessments. The desired outcomes are expressed as the course learning or educational objectives. Ralph Tyler (1949) explains this rational:

> Educational objectives become the criteria by which materials are selected, content is outlined, instructional procedures are developed, and tests and examinations are prepared. ... The purpose of a statement of objective is to indicate the kinds of changes in the student to be brought about so that instructional activities can be planned and developed in a way likely to attain these objectives (pp. 1, 45).

Backward design suggests instructors do not design the assessments at the end of the course (Wiggins & McTighe, 1998). Assessments should measure students' achievements in the course learning objectives. Evidence of what achievement means should be expressed in the course learning objectives at the beginning of the semester. Thus, learning objectives not only describe what students will be able to do at the end of the semester but also express the kinds of evidence that demonstrate those abilities. This brings transparency for the students; at the beginning of the semester they know what they will learn and what evidence is needed to demonstrate what they have learned.
Standards-based grading, which is based on backward design, was first introduced in the United States public K-12 education in the 1990s. In that system, academic standards for what students should know and be able to do were developed (Marzano, 2010; Reeves, 2002). After more than two decades, some universities, including Purdue University, have started to change their grading system from traditional summative score-based grading to standards-based grading.

While scholars use different terminologies to refer to standards-based grading (e.g., objective-driven, criteria-based, competency-based), they use it to describe similar (but not always the same) method of assessment. For example competency-based grading, unlike standards-based grading, is based on mastery of units or modules and students can earn credit or advance in learning content at their own pace. However, all of these methods have the same underlying concept, which is assessing students’ learning based on well-articulated learning objectives, and can be considered the same for the purpose of this dissertation. A detailed review of different grading systems is provided by Sadler (2005).

Heywood (2014) defines standards-based grading as “the measurement of the quality of students’ proficiency towards achieving well defined course objectives” (p. 1514). Similarly, Sadler (2005) requires that grades “represent how well students achieve the course objectives” (p. 179). Standards-based grading is criterion-referenced not norm-referenced. In other words, students are graded based on competency or what they can do, regardless of how other students in the
course perform on the same assigned task (Carberry et al., 2012; Heywood, 2014; Sadler, 2005). In a standards-based grading system, the course assessments are directly connected to the course learning objectives and are not a series of separate course assignments (Carberry et al., 2012).

Standards-based grading provides educational advantages for students. Because standards-based grading assesses students' competency of the course learning objectives, it provides clear, meaningful, and personalized feedback for students related to achievement of the course learning objectives and helps them identify their weaknesses in the course (Atwood & Siniawski, 2014). In addition, because the students are aware of the course learning objectives and a student's grade is independent of other students' performance, it provides "fairness and transparency" (Sadler, 2005).

Some disciplines such as medicine and some professions such as airline pilots have used competency-based testing (Heywood, 2014). Also in the liberal arts some universities, such as Alverno College, have utilized learning systems that assess competencies (or as they call them, abilities) (Mentkowski et al., 2000). Use of standards-based grading is relatively new in engineering. Carberry et al. (2012) piloted five STEM courses with a standards-based grading system in order to investigate its affective and cognitive influence on students' behaviors. The results suggest that a standards-based grading system was perceived as valuable, increased students' domain-specific self-efficacy, and helped students develop more sophisticated beliefs about STEM knowledge (Carberry et al., 2012).
2.2 Learning analytics and predictive modeling

Learning analytics is a new field of inquiry that uses data-driven decision-making techniques to support teaching and learning (van Barneveld, Arnold, & Campbell, 2012). In this section, I briefly explain learning analytics and its related concepts.

The goal of analytics is to summarize the relationships in data and make it understandable for decision-making (Witten & Frank, 2005). The relationships in data are summarized in either patterns or models. A pattern is a local relation in data; a model is a global relation in data (Hand, Mannila, & Smyth, 2001). The techniques used to find the relationships in data are called modeling techniques.

In general, there are two types of modeling techniques: predictive modeling and descriptive modeling (Hand et al., 2001). The main difference between predictive modeling and descriptive modeling is the purpose of the modeling. The purpose of descriptive modeling is to make sense of (large) data, make it understandable, and describe relationships in it (Hand et al., 2001). In other words, descriptive modeling creates an understandable summary of data. A very simple example of descriptive modeling is basic descriptive statistics such as calculating an average or creating a histogram of the distribution of the data. Visualization techniques are typically descriptive; they usually summarize a lot of information in a way that can be understood quickly.

Predictive modeling uses current data to create a model that predicts a specific (future) characteristic of an object (e.g., a student) based on other observed
characteristics (Hand et al., 2001). Creating a predictive model has two phases, training and testing. In the training phase, a model is created based on (part of) the current data using predictive modeling techniques. The data that is used to create (i.e., train) the model is called training data. Then in the testing phase, the accuracy of the model is evaluated based on (another part of) the current data. The data that is used to test the model is called test data. Ideally the training dataset and the test dataset are independent. Because the model is built based on the training data, it should be tested based on another dataset. Only in this case we can truly test the model for predicting future data.

Predictive modeling itself has two types: predictive modeling for classification and predictive modeling for regression (Hand et al., 2001), which are based on the type of the predicted variable (e.g., grade). Predictive modeling for classification is used when the type of predicted variable is categorical (e.g., predicting a letter grade), and predictive modeling for regression is used when the type of predicted variable is continuous (e.g., predicting a grade point).

### 2.3 Predict students' success in a course

The first step in helping students who may fail a course is to identify them early in the semester. The course instructor can use this information to help at-risk students be successful in the course. In an attempt to predict students' grades early in the semester, some instructors use course syllabus grading criteria and apply it to the available performance information to calculate an early grade for students. This method cannot be used at the very beginning of the semester (e.g., week 4) when the
majority of students’ performance information is not available (e.g., exam and project scores). It can also be extremely inaccurate at the beginning of the semester, because most of the performance data is not available. This method can only be accurate closer to the end of the semester when the majority of the performance information is available. Using such inaccurate methods at the beginning of semester can result in wrong predictions and lead to student mistrust in the predictions. In contrast, predictive modeling maps students’ success at the end of the semester with the performance data during the semester and weights different grades based on their predictive power.

2.4 Generic early warning systems

With the use of predictive modeling techniques, it is possible to predict students’ success in a course and identify at-risk students (Lackey, Lackey, Grady, & Davis, 2003; Olani, 2009). A predictive model can be used as an early warning system, which identifies at-risk students in courses and informs both the instructor and the students of their performance. Instructors can use different strategies to communicate to the at-risk students and provide them guidelines and support for improving their performance in the course. Use of an early warning system in a course, along with guidelines and support, can increase students’ success in a course (Arnold & Pistilli, 2012; Essa & Ayad, 2012; Macfadyen & Dawson, 2010).
2.4.1 Examples of existing early warning systems

Course Signals (Arnold & Pistilli, 2012) is a pioneer early warning system that is currently being used at Purdue University. The Course Signals' predication model is based on a model created by Campbell (2007) employing a regression model. This model utilizes campus-wide access data, i.e., log data from a Course Management System (CMS), from 27,276 students that enrolled in 608 courses in 75 departments across nine colleges, without differentiating courses, departments, or even colleges; it uses the same model for all courses across the campus. The prediction algorithm has four main input components to predict students' grades in each course: performance in the course so far (e.g., quizzes, homeworks, exams), effort or interactions with the CMS (e.g., access to course materials) compared to other students' in the course, prior academic history (e.g., academic preparation, high school GPA, and SAT/ACT), and students' characteristics (e.g., residency, age, and credits attempted) (Arnold & Pistilli, 2012; Pistilli & Arnold, 2010).

According to its creators (Pistilli & Arnold, 2010), Signals is a behavior-based model:

This means that while some consideration is given to data such as past academic performance, far more weight is given to student effort in terms of interaction with the CMS (e.g., taking online quizzes, reading articles, and exploring resources) and help-seeking behavior, so that students start on an even playing field at the beginning of the course (p. 23).
Based on Campbell (2007) and Course Signals related research, Macfadyen and Dawson (2010) used CMS data to create another early warning system. They used regression and CMS access/log data of an online course with 118 students to build an early warning system and predict students’ grade in the online course. From 17 variables used in a multiple linear regression model, only three of them (total number of discussion messages posted, number of assessments finished, and number of mail messages sent) were significant predictors of students’ success, explaining only 33% of variability in data ($r^2=0.33$). To predict at-risk students, a binary logistic regression was used, and the same three variables were significant in predicting at-risk students. Overall accuracy of this model was 73.7%. Type II error (failure to identify at-risk students) was 12.7%.

The Student Success System (S3) by the Desire2Learn company is another early warning system that uses predictive modeling (Essa & Ayad, 2012). S3 designers argue they employ a “highly-generalizable modeling strategy” using an ensemble modeling method. To create the ensemble model, they combined four different prediction models with four different sets of predicting variables: attendance, competition, participation, and social learning. For each of these domains, a different predictive modeling system is used and the final decision (successful or not) is made by combining results of the four models. S3 uses support vector machine (SVM) for the predictions. I did not find any assessment of the accuracy of this system.

---

1 Based on a conversation with Desire2Learn senior research scientist at LAK 2014 conference.
Personalized grade prediction advisor (pGPA) (Sheehan & Park, 2012) uses students’ academic history (e.g., grades) along with temporal (e.g., when a grade has been earned) and contextual (e.g., how many credit hours the student is carrying each semester) information to predict students’ success in a course. pGPA, similar to other prediction systems, uses the same prediction model for different courses. I did not find any assessment of the accuracy of this system. With only using past data, ignoring students’ current performance, and without reporting the system’s accuracy, pGPA is likely to stigmatize students rather than helping them succeed. The shortcomings of these early warning systems are described in the next section in more details.

2.4.2 Shortcomings of early warning systems

One common and major problem with early warning systems that are currently being used is that they typically employ a general model that cannot address the complexity of all courses. Over the past decade, more instructors use active learning strategies (e.g., flipped classroom, group discussions) instead of the traditional extensive lecturing (Hurtado, Eagan, Pryor, Whang, & Tran, 2012). The new pedagogies are usually implemented along with new assessment methods such as team-based projects that do not fit in the traditional homework and exam assessment framework. Thus, the course syllabus and the assessment components can vary largely from one course to another. Therefore, using one model for different courses can significantly reduce the accuracy of predictions of students’ academic success.
While S3 designers critique Course Signals for using only one model, they use the same four models for different courses. While this approach (using the same model for all courses) is cost effective for higher education institutions, it lowers the accuracy of the model. Creating predictive models at the course level increases the accuracy of the model (Macfadyen & Dawson, 2010), but with higher cost. Also, there needs to be enough data from each course to train and test course-specific models.

To date, all early warning systems have been developed based on score-based grading. In a standards-based course, because assessments are designed based on the learning objectives, it may be possible to find relationships between learning objective achievement and assessment grades. Finding these relations enables us to predict a student’s grade on a later course assessment based on early achievement of the learning objectives. These relations can be used to create a prediction model, which can be utilized as an early warning system. In addition, because in standards-based grading each assessment is graded based on several learning objectives, it provides more data points to use in the prediction models.

Most early warning systems rely heavily on access data and not performance data. While this approach might be effective for online learning courses, in which students have a lot of interactions with the CMS, it can be problematic for many face-to-face courses in which most of the learning activities and evidence of learning is happening offline. First of all, from a practical point of view, many face-to-face courses do not use a CMS for the course or for some of the course activities. For
example, many courses at Purdue University now use Piazza for discussions/questions instead of Blackboard, so a Blackboard-based early warning system has no access to class data. Secondly, in face-to-face classes, students’ effort in the course cannot be measured by analyzing only online CMS access data. Third, students’ engagement in online activities measured by CMS access data may not be related to their course grade (Beer, Jones, & Clark, 2009) and can be influenced by a number of factors including but not limited to teacher participation, course design, class size, student gender, and student age (Beer, Clark, & Jones, 2010).

2.5 Course specific prediction systems

An extensive amount of research has been conducted to determine which factors correlate and/or predict students’ academic success in a course. In this section, I review some of the studies that have been conducted to predict students’ grades in STEM related courses.

I searched the terms “student” and “success” on paper titles in the engineering village database (including Compendex and Inspec) to find STEM related papers that discussed students’ success. This database includes the Journal of Engineering Education, American Society for Engineering Education (ASEE) annual conference proceedings, Frontiers in Education (FIE) conference proceedings, The International Conference of Computer-Supported Collaborative Learning (CSCL) proceedings, The International Conference of Learning Sciences (ICLS) proceedings, some of the Association for Computing Machinery (ACM) publications, and a number of other engineering/education related peer-reviewed
journals and conferences. From 691 finding, I selected 180 based on their title and their abstracts. Only 18 studies (listed in Table 2.1) reported factors related to students' success in a specific STEM course. Table 2.1 is a list of these studies along with some basic information about each study.

As is illustrated in Table 2.1, the majority of these studies focused on predicting students' grade in a course at the end of the semester using academic factors available before the start of the semester (e.g., cumulative GPA, grade in a prerequisite course) and non-academic factors (e.g., gender, age). Neither course instructor nor students can influence what has happened in the past before taking the course (e.g., student GPA, previous grades) or non-academic factors (e.g., gender, race, socio-economical status). In addition, the course instructor typically does not have access to students' prior records to use for prediction purposes, unless through a CMS that keeps this information blind.

Surprisingly, few of these studies have utilized the academic performance data available during the semester, which logically can be the best predictor of the course grade. From the studies listed in Table 2.1, only two used course performance in the prediction model. One of these studies (Huang & Fang, 2013) compared four different methods to predict students' grades in an engineering dynamics course using 323 students' data from four semesters. In this study, three mid-term scores were used as indicators of students' performance during the semester. In addition, cumulative GPA, Statics grade (a pre-requisite course), Calculus I and II grades, and Physics grade were used as indicators of students'
Table 2.1 - STEM course specific prediction studies.

<table>
<thead>
<tr>
<th>Study (Author, Year)</th>
<th>Participants</th>
<th>Method</th>
<th>Related Factors</th>
<th>Unrelated Factors</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Hostetler, 1983)</td>
<td>79</td>
<td>Programming Multiple Regression</td>
<td>GPA</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Diagramming</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Reasoning tests</td>
<td></td>
</tr>
<tr>
<td>(Corman, 1986)</td>
<td>83</td>
<td>Programming Marketing Regression</td>
<td>Major GPA</td>
<td>Cognitive style</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Gender (only for marketing)</td>
<td>Personality type</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Prev. programming courses</td>
<td>Learning ability</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Age (for programming class)</td>
<td></td>
</tr>
<tr>
<td>(Felder, Forrest, Baker-Ward, Dietz, &amp; Mohr, 1993)</td>
<td>123</td>
<td>Chemical Engineering Correlation, Multiple regression</td>
<td>Predictors: GPA, Prev. chemistry and physics courses</td>
<td>---</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Correlations: college admission criteria, freshman year grades, and scores on the Learning and Study Strategies Inventory</td>
<td>---</td>
</tr>
<tr>
<td>(Clark &amp; Riley, 2001)</td>
<td>407</td>
<td>Chemistry 2 factor ANOVA</td>
<td>Personality type</td>
<td>---</td>
</tr>
<tr>
<td>(Devens &amp; Walker, 2001)</td>
<td>3087</td>
<td>Intro to engineering Visualization Fit line (similar to regression)</td>
<td>SAT (correlation but not prediction)</td>
<td>Gender</td>
</tr>
<tr>
<td>(Diefes-Dux, 2002)</td>
<td>404</td>
<td>Math Descriptive statistics</td>
<td></td>
<td>Math bridge program</td>
</tr>
<tr>
<td>(Cummings, Lockwood, &amp; Marx, 2004)</td>
<td>275</td>
<td>Physics Correlation</td>
<td>Aptitude toward problem solving</td>
<td>---</td>
</tr>
<tr>
<td>(Carpenter &amp; Hanna, 2007)</td>
<td>260</td>
<td>Calculus I Multiple linear regression</td>
<td>Math ACT score</td>
<td>---</td>
</tr>
<tr>
<td>(Doyle, Kasturiratna, Richardson, &amp; Soled, 2009)</td>
<td>270</td>
<td>Programming Pearson’s chi-square</td>
<td>Hours spent using online math tutorial program</td>
<td>---</td>
</tr>
<tr>
<td>(Baxter, Hungerford, &amp; Helms, 2011)</td>
<td>210</td>
<td>Programming Multiple linear regression</td>
<td>ACT math score</td>
<td>Gender</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Complete a pre-req. programming course</td>
<td>Major</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Student aptitude toward tech.</td>
<td>Student</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Participate in optional lab section</td>
<td>Prior</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>High school and prior college computer training</td>
<td>training</td>
</tr>
<tr>
<td>Study (Author, Year)</td>
<td>Participants</td>
<td>Method</td>
<td>Related Factors</td>
<td>Unrelated Factors</td>
</tr>
<tr>
<td>----------------------</td>
<td>--------------</td>
<td>--------</td>
<td>-----------------</td>
<td>-------------------</td>
</tr>
</tbody>
</table>
| (Harris, Harris, & Lambert, 2011) | 204 --- | Computers in business | Hierarchical moderated regression analysis | Age  
Computer self-efficacy  
Computer anxiety  
Conscientiousness (motivation)  
Gender  
Hours worked per week  
Class hours per semester  
Work experience  
Need for achievement |
| (Durfee, Loendorf, Richter, Geyer, & Munson, 2012) | 1600 1st, 2nd, 3rd | 8 engineering courses | Descriptive statistics | Attendance |
| (Huang & Fang, 2013) | 323 --- | Dynamics | Multiple linear regression, Neural Networks, SVM | GPA  
Previous relates courses grades  
Midterm scores |
| (Bahovec, Erjavec, & Čičmešja, 2013) | 107 --- | Statistics | Logistic regression | Attendance  
Math score  
Informatics grade |
| (Olama, Thakur, McNair, & Sukumar, 2014) | 11,000 --- | 270 math courses | Logistic regression  
Neural network | Gender  
Age  
Employed |
| (Porter, Zingaro, & Lister, 2014) | --- 1st | Computer Science | Correlations | Homeworks  
Quiz  
Participation in online discussions |
| (Simpson & Fernandez, 2014) | 53 1st, 2nd, 3rd | ECE courses | Linear regression  
Hierarchical Linear Model | Clicker questions  
Prerequisite math and physics grade |
| (Sinapuelas& Stacy, 2015) | 61 1st | Chemistry | Hierarchical Linear Model | Learning approach |

Table 2.1 continued - STEM course specific prediction studies.
performance before starting the course. Students’ grades at the end of the semester were predicted using multiple linear regression (MLR), multilayer perceptron (MLP) neural network, radial basis function (RBF) neural network, and support vector machine (SVM).

Table 2.2 illustrates the accuracy of these methods using different variables. As expected, model 4, which employs all available data, has the most accuracy among the different models. In addition, SVM is the most accurate prediction method (in most cases). However, a comparison of models 1 and 2 reveals that adding grades of previous related courses to GPA does not result in a significant increase in the accuracy of the model. Furthermore, using only the first mid-term exam with an accurate prediction method (e.g., SVM) yields a similar result to using pre-course performance such as GPA or previous courses grades. These results clearly demonstrate the value of performance data during the semester for predictive purposes. Adding other performance data such as homework and quiz scores can also increase the accuracy of the models.

Table 2.2 - Accuracy of the predictions using different prediction methods and variables (Huang & Fang, 2013)

<table>
<thead>
<tr>
<th>#</th>
<th>Variables used in the model</th>
<th>MLR</th>
<th>MLP</th>
<th>RBF</th>
<th>SVM</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Only GPA</td>
<td>53.8%</td>
<td>53.4%</td>
<td>53.9%</td>
<td>53.7%</td>
</tr>
<tr>
<td>2</td>
<td>GPA, statics, calculus I &amp; II, and physics grades</td>
<td>54.9%</td>
<td>52.9%</td>
<td>53.3%</td>
<td>54.2%</td>
</tr>
<tr>
<td>3</td>
<td>1st mid-term score</td>
<td>49.0%</td>
<td>48.9%</td>
<td>51.5%</td>
<td>52.5%</td>
</tr>
<tr>
<td>4</td>
<td>All variables (2nd model + 3 mid-term scores)</td>
<td>61.3%</td>
<td>59.5%</td>
<td>62.9%</td>
<td>64.0%</td>
</tr>
</tbody>
</table>

In another study Olama and his colleagues (2014) used in-semester performance data to predict more than 11,000 students’ success/failure in 270 mathematical courses. They used homework and quiz grades as well as
participation in online discussions as predicting variables. A neural network model predicted students’ success/failure with 84% accuracy at week 3. Logistics regression was used to compare the relative importance of predictive variables. Homework grades were stronger predictor of success/failure in the course than quiz grades and participation in discussions. Similar to previous studies, these results clearly demonstrate the value of performance data during the semester for predictive students’ success in a course. In addition, unlike mid-term exams, homework starts earlier in the semester and using these grades as predictors of success result in more accurate predictions early in the semester.

2.6 Significance of this research

First and foremost, this study is the first attempt to create predictive models based on achievement of the course learning objectives and provide recommendations for creating learning objective based course specific prediction models to identify at-risk students. In addition to the described benefits of early warning systems, an objective-based prediction model has educational benefits for the students and instructional advantages for the course designers and instructors. Such a prediction model creates a meaningful and clear connection between achievement of the learning objectives and success in the course. This helps both instructors and students understand which learning objectives are important and lead to success or failure in the course. From an instructional design perspective, the model highlights the learning objectives for a course that are important for passing or failing the course. The learning objectives highlighted by the model may not be
the same as the ones the course designer or the instructor intended to be the most important learning objectives.

In this study, the prediction models were optimized by prioritizing students who failed the course over those who passed in order to increase the identification of at-risk students. The existing early warning systems employ one model for predicting students’ success in all courses, which decreases accuracy of the predictions. The prediction models that were developed in this study are course-specific to increase the accuracy of the predictions. In addition, this study identified the minimum number of students in a course that accurately train the prediction models.

The selected course for this dissertation is a more complex course compared to most courses for which predictive models have been developed. Because of the variety of assessment methods used in this course, it is not possible to use the current early warning systems such as Course Signals for this course. This dissertation research can be used as a prototype to move toward course-specific systems, especially for high enrollment and complex courses. The techniques and procedures that were developed in this dissertation can be adapted for other courses that utilize standards-based grading.
CHAPTER 3 - METHODS

This chapter described the data source for this study including the course and its settings, and the data that were used in the analysis. I also described the prediction models and how they were trained, verified, and tested, and the feature selection methods that were used to improve the models.

3.1 Data source and settings

This study used secondary data collected during the Spring 2013 and 2014 semester offering of a first-year engineering (FYE) course at a large Midwestern U.S. university. In each semester, approximately 1650 FYE students enrolled in the course. Nearly 20% of the FYE students were female and about 20% were international students. This course is a required second semester, 2-credit hour course for all FYE students. In this course, students learn how to make evidence-based engineering decisions in diverse teams, develop problem-solving, modeling, and design skills that they will use as an engineer, learn how to use computer tools to solve fundamental engineering problems, and continue to improve their teaming and communication skills.

In this FYE course, homeworks were graded based on learning objectives. Other course assessments including quizzes and exams were designed based on the learning objectives but graded with a traditional score point system. Each
homework was designed and assessed based on 6-7 learning objectives (e.g., draw
and interpret flowcharts containing decision branches to characterize an
engineering problem; write evidence-based rationales). Learning objectives were
assessed on a four level scale: no evidence (0), under achieved (1), partially
achieved (2), or fully achieved (3).

The course learning objectives were categorized into six topics and four
types. The six topics were: Linearization and Trend Lines, User Interface, Logic and
Conditional Statements, Displaying Data and Statistics, Foundations of
Programming, and Building/Using Loops. Course learning objective types were:
Problem Interpretation, Non-computational Application, Computational Application,
Results Interpretation/Use (Figure 3.1).

This course is a good example to showcase how to create and use predictive
models for a course. First, it enrolls a large number of students, so not only has it
enough data to train the prediction models, but also it is possible to train the
prediction models on different subsets of the data to investigate the effect of sample
size on the models. Some of the models require a fairly large sample size to train
(Hand et al., 2001). Second, it involves several different graded components (e.g.,
quizzes, homeworks, exams, projects). Thus it is possible to investigate which
assessments have greater prediction power. Third, most of the assessments are
collected on a weekly basis. Thus, this course generates a large number of number of
performance data each week. Fourth, some of the assessments (i.e., homeworks) are
graded based on learning objectives. As such, all graders in the course (instructors,
Figure 3.1 – ENGR 132 concept map of learning objectives (Hylton & Diefes-Dux, unpublished).
graduate and undergraduate teaching assistants) are grading based on well-defined rubrics. Thus, it is more likely that the grades are reliable. In addition, it is possible to link achievement on the learning objectives to the success in the course.

3.2 Data

In-semester performance data from the Spring 2013 and Spring 2014 offerings of the course were used in this study. The in-semester student performance data for this course included grades for attendance, quizzes, and weekly homework as well as team participation, project milestones, mathematical modeling activity tasks, and exams (Table 3.1).

<table>
<thead>
<tr>
<th>Assessment Component</th>
<th>Standards-based Grading</th>
<th>Percentage Sp13/Sp14</th>
<th># of data points per student at week 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>In-class Participation/Quizzes</td>
<td>No</td>
<td>5%</td>
<td>10</td>
</tr>
<tr>
<td>Team Participation (TP)</td>
<td>No</td>
<td>5%</td>
<td>0</td>
</tr>
<tr>
<td>Homework (HW)</td>
<td>Yes</td>
<td>15%</td>
<td>33</td>
</tr>
<tr>
<td>Design Project</td>
<td>No</td>
<td>20%</td>
<td>0</td>
</tr>
<tr>
<td>Mathematical Modeling Activity</td>
<td>No</td>
<td>15%</td>
<td>0</td>
</tr>
<tr>
<td>Written Exams</td>
<td>No</td>
<td>40%</td>
<td>1</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td></td>
<td>44</td>
</tr>
</tbody>
</table>

In this study, performance data available at the end of week 5 of the semester, which included the homework learning-objectives scores and grades for quizzes and written exam 1, referred to in this dissertation as midterm exam, were used. By the end of week 5, students had participated in one written exam, 10 quizzes, and five homeworks with 33 learning objectives. This may be enough data to predict the students’ success/failure in the course while they still have enough time to learn of the course material to successfully pass the course.
Homeworks were assigned to students for each week based on the week’s topic. Students had about one week to solve the problems in the homework individually and submit their responses. Each homework was designed and graded based on 6-7 learning objectives on a zero (no evidence) to three scale (fully achieved). Quizzes were administrated in every class session. Quizzes were graded based on a zero to 10 grade point scale. Midterm exam questions were designed based on learning objectives but graded based on a zero to 100 grade point scale. Students were asked to solve the midterm exam problems during a 60-minute exam session individually.

There were some missing data in the datasets. In the course, if a student missed a homework or quiz, she received a zero grade for that assessment. Thus, all missing data were replaced by zero. Table 3.2 shows the number of missing cases for each assessment.

Table 3.2 – Number of missing grades. Missing grades replaced by zero.

<table>
<thead>
<tr>
<th>Semester</th>
<th>H1*</th>
<th>H2</th>
<th>H3</th>
<th>H4</th>
<th>H5</th>
<th>Q1** a/b</th>
<th>Q2 a/b</th>
<th>Q3 a/b</th>
<th>Q4 a/b</th>
<th>Q5 a/b</th>
<th>Midterm Exam</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sp 2013</td>
<td>85</td>
<td>63</td>
<td>172</td>
<td>127</td>
<td>143</td>
<td>40</td>
<td>40</td>
<td>40</td>
<td>40</td>
<td>41/40</td>
<td>40</td>
</tr>
<tr>
<td>(N=1560)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sp 2014</td>
<td>131</td>
<td>255</td>
<td>132</td>
<td>241</td>
<td>169</td>
<td>46</td>
<td>46</td>
<td>46</td>
<td>46</td>
<td>46</td>
<td>46</td>
</tr>
<tr>
<td>(N=1413)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

* H1-H5 are homeworks for weeks 1-5.
** Q1-Q5(a/b) are quizzes for weeks 1-5; each week 2 quizzes were administrated.

### 3.3 What is success in the course?

Although passing a course is usually considered as receiving a grade of D or better, students who barely pass a course with a D letter grade can also benefit from early support and intervention in the course. For this course, at least a C grade is
required to complete the first-year engineering program and matriculate to an engineering disciplinary program. In addition, most engineering programs have a minimum requirement of maintaining a C or better GPA to stay in the program. For example, at Purdue University, a student is placed on probation if her GPA is less than 2.0 (Purdue University, 2012), which is equivalent to a C letter grade. Furthermore, if a student decides to transfer to another school or program, she typically needs to pass most of the courses with grades of A or B, and sometimes C grades. For example, most of the engineering programs at Purdue University have a minimum GPA requirement of 2.5 for transfer students (Purdue University, 2013). For these reasons, in this study, similar to other studies (e.g., Macfadyen & Dawson, 2010), success was defined as earning at least a C grade, which was equivalent to a final grade of 68% or higher in the course. D, F, and W (withdraw) grades were defined to be failing grades for the course. Figure 3.2 shows the distribution of letter grades for students who enrolled in the FYE course in the Spring 2013 and 2014 semesters. Because some students withdrew from the course after the first few weeks of the semester and there was not a clear record of these students in the course grade book, students who were not issued a final grade, or had a zero final grade and missed the midterm exam were assumed to have withdrawn from the course.
Figure 3.2 – Distribution of letter grades for FYE course in the Spring 2013 and 2014 semesters.

3.4 Prediction modeling methods

Six different prediction modeling methods that are commonly used in educational data mining (Baker & Yacef, 2009; Romero & Ventura, 2010) were chosen to build prediction models to identify at-risk students. In addition to these six models, the results led to creating a seventh ensemble model. These modeling methods are described below.

3.4.1 Logistic Regression

Logistic Regression (Log Reg) is a reliable prediction method commonly used in educational settings (e.g., Braunstein et al., 2008; Eckles & Stradley, 2012; Hendel, 2007). It calculates the probability of a categorical variable (e.g., letter grade, pass/no-pass) from a number of predicting variables (Kutner, Nachtsheim, Neter, & Li, 2005). Eq. 1 shows a logistic regression with $m$ predictor variables (e.g., scores on homework learning objectives, midterm exam, etc.) and one outcome variable $Y$ (e.g., probability of passing the course):
\[ Y = \ln \left( \frac{p}{1-p} \right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_m x_m + e \]  \quad (Eq. 1)

where:

- \( p \) is the probability of the desired outcome (e.g., passing the course) based on predictors \( x_1 \) to \( x_m \) (e.g., quiz grade), and
- \( \beta \) coefficients are estimated in the training phase based on the training data.

### 3.4.2 Support Vector Machine

Support Vector Machine (SVM) finds a hyperplane (e.g., a line in the 2D space) that separates two categories of data (Cortes & Vapnik, 1995). Finding the hyperplane with the maximum margin from both categories (e.g., fail/pass student categories) is an optimization problem. SVM is only sensitive to the data points close to the border of two categories. If the two categories are not linearly separable, non-linear SVM can be used to find an optimum surface. A linear hyperplane can be describe as:

\[ w \cdot X + b = 0 \]  \quad (Eq. 2)

where:

- \( w \) is the normal to the hyperplane,
- \( X \) is the vectors of predictor variables \( x_1 \) to \( x_m \) and
- \( \frac{b}{\|w\|} \) is the perpendicular distance from the hyperplane to the origin.

To find the hyperplane with the maximum margin from both categories a Lagrange objective function (Eq. 3) is minimized with respect to \( w \) and \( b \) values:

\[ L_p = \frac{1}{2} \|w\|^2 - \sum_{i=1}^{n} \alpha_i y(i)[x(i) \cdot w + b] + \sum_{i=1}^{n} \alpha_i \quad \alpha_i > 0 \quad \forall \quad (Eq. 3) \]
where:

\[ n \] is the number of cases (e.g., students), and

\[ y(i) \] is the outcome for case \( i \) (e.g., student \( i \) passed or failed)

3.4.3 Decision or Classification Tree

Decision or Classification Tree (DT) is a modeling method based on partitioning (Maimon & Rokach, 2008). In each step, it partitions the data based on one variable (e.g., midterm exam grade) until all data in each node have only one category label (e.g., pass or fail) or all variables have been used (Hand et al., 2001). Partitioning is done by defining a score function that calculates the purity of all possible nodes and selects the variable that generates the purest nodes. This study used Gini gain (Breiman, Friedman, Olshen, & Stone, 1984) as the score function for training the tree method, which is described in more detail in section 3.6.2.

3.4.4 Multi-Layer Perceptron

Multi-Layer Perceptron (MLP) is an Artificial Neural Network (ANN). In general, ANNs try to mimic the brain structure. An ANN is a network of neurons (i.e., nodes) that are connected together with different weights. MLPs can have multiple input variables (input layer) and one or more hidden layers with different numbers of nodes. Depending on the type of network, there may be one or more outputs (Hand et al., 2001). Eq. 4 illustrates a MLP with \( m \) input variables and a binary output \( y \).

\[
y = \text{sign}[f(x)] \quad \text{where} \quad f(x) = \sum_{i=1}^{m} w_i x_i + b \quad (\text{Eq. 4})
\]
where:

\[ y \text{ is the predicted class (e.g., pass or fail the course) based on predictors } x_1 \text{ to } x_{nu}, \]

and

\[ w_i \text{ and } b \text{ are calculated during the training phase.} \]

### 3.4.5 Naive Bayes Classifier

Naive Bayes Classifier (NBC) is a simple probabilistic classifier that calculates a conditional probability distribution over the output of a function based on applying Bayes’ theorem (Stuart & Ord, 1994) with the (naive) assumption of independence between the predictive variables (Russell & Norvig, 1995). Although this assumption is often violated (e.g., the midterm exam and quiz grades are not independent, see chapter 5 for correlations between the variables), the NBC performance can be comparable to more advanced methods such as SVM (Rennie, Shih, Teevan, & Karger, 2003). Eq. 5 shows the Bayes conditional probability rule and the naive assumption.

\[ P(C|X) = \frac{P(X|C)P(C)}{P(X)} \quad \text{Naive Assumption: } P(X|C) = \prod_{i=1}^{m} P(x_i|C) \quad (Eq. 5) \]

where:

- \( P(C|X) \) is the posterior probability of class \( C \) (e.g., pass or fail the course) given predictors \( X \) (e.g., quiz and homework grades),

- \( P(C) \) is the prior probability of class,

- \( P(x_i|C) \) is the likelihood which is the probability of predictor given class, and

- \( P(X) \) is the prior probability of predictor.

A-priori probabilities are calculated based on the training dataset.
3.4.6 K-Nearest Neighbor

K-Nearest Neighbor (KNN) is a non-parametric classifier. Unlike the methods described above, it does not train a model with parameters. KNN classifies an object (e.g., a student) by a majority vote of its K neighbors (Friedman, Bentley, & Finkel, 1977). Thus, instead of model parameters, it only calculates the distance between the objects. In this study, the five nearest neighbors were used to identify at-risk students. The euclidian distance between two points $(x$ and $y)$ was calculated to find the nearest neighbors (Eq. 6).

$$d_E(x, y) = \sum_{i=1}^{m} \sqrt{x_i^2 - y_i^2} \quad \text{(Eq. 6)}$$

3.4.7 Ensemble method

An Ensemble model is created by training more than one model on the dataset and combining them during prediction based on a majority vote of the models (Rokach, 2010). One of the main advantages of using an Ensemble method is that it is not necessary to decide a-priori which model to use and it is possible to use a combination of multiple models. In this study, an Ensemble model consisting of three models was used (See section 4.2 for more details).

3.5 Train, verify, and test the models

For all of these models, the same data were used to train, verify, and test the models (Table 3.3). Spring 2013 data were used to train, verify, and optimize the prediction models. The students’ performance data (i.e., achievement of learning objectives and course grades) from Spring 2013 were randomly (with control for pass/fail ratio) divided into three datasets: 50% for training, referred to as the train
dataset; 25% for comparing different methods and tuning the models, referred to as verify 1 dataset; and 25% for the secondary verification after improving the models, referred to as verify 2 dataset. Spring 2014 data were used to test the models.

<table>
<thead>
<tr>
<th>Semester</th>
<th>Dataset</th>
<th>Total # of students</th>
<th># of passed students</th>
<th># of failed students</th>
</tr>
</thead>
<tbody>
<tr>
<td>Spring 2013</td>
<td>Train</td>
<td>780</td>
<td>723</td>
<td>57</td>
</tr>
<tr>
<td>Spring 2013</td>
<td>Verify 1</td>
<td>390</td>
<td>361</td>
<td>29</td>
</tr>
<tr>
<td>Spring 2013</td>
<td>Verify 2</td>
<td>390</td>
<td>361</td>
<td>29</td>
</tr>
<tr>
<td>Spring 2014</td>
<td>Test</td>
<td>1413</td>
<td>1266</td>
<td>147</td>
</tr>
</tbody>
</table>

While the number of students who failed the course is less than or close to 10%, it is critical that the models identify these students correctly. As the goal of this study is to identify at-risk students, it is important to achieve high predictive accuracy for the students who failed the course. Therefore, in addition to the overall accuracy of the models (Eq. 7), the accuracy for the students who passed (Eq. 8) and failed (Eq. 9) the course were also calculated. In order to compare the models, a $F_{1.5}$ score (Eq. 10) was also calculated. The $F_{1.5}$ score is a harmonic mean of accuracy for the student who passed and failed the course that weights the accuracy for students who failed more than for students who passed (van Rijsbergen, 1979). For the models, a negative result means the student is not at-risk and will pass the course; a positive result means the student is at-risk and will not pass the course.

\[
\text{Accuracy} = \frac{TN + TP}{\text{Total number of students}} \quad (\text{Eq. 7})
\]

\[
\text{Accuracy (Pass)} = \frac{TN}{\text{Number of passed students}} = \frac{TN}{TN + FP} \quad (\text{Eq. 8})
\]

\[
\text{Accuracy (Fail)} = \frac{TP}{\text{Number of failed students}} = \frac{TP}{TP + FN} \quad (\text{Eq. 9})
\]
\[ F_{1.5} = \frac{(1 + 1.5^2).TP}{(1 + 1.5^2).TP_s + 1.5^2.FN + FP} \]  
(Eq. 10)

where:

* **TP or True Positives** is the number of students who failed and were identified as at-risk.

* **TN or True Negatives** is the number of students who passed and were not identified as at-risk.

* **FN or False Negatives** (type II error) is the number of students who failed the course but were not identified by the models as at-risk.

* **FP or False Positives** (type I error) is the number of students who passed the course but were identified by the models as at-risk.

### 3.6 Feature Selection

Feature selection is the process of selecting a subset of features (i.e., variables) to use in the training of a model to improve the results (James, Witten, Hastie, & Tibshirani, 2013). Feature selection methods try to select variables that have more predictive power and are more related to the predicted variable and filter the ones that are not useful in predictions. In the prediction of at-risk students in this study, using a subset of the variables (i.e., grades) may yield a more generalizable model than using all of them. Selecting a subset of variables that are more important in students’ success is also valuable from an instructional point of view. Especially in the case of learning objectives, feature selection highlights the possible *threshold* learning objectives that are important in understating other
learning objectives and being successful in the course (Meyer & Land, 2013). The three different feature selection methods used in this study are described below.

3.6.1 Correlations

One way to select a subset of predicting variables that are more related to the predicted variable is by calculating the correlations between them. Figure 3.2 shows the correlation of the 44 predicting variables with success in the course. Variables are divided into four clusters. Depending on the researcher/instructor approach, this graph can provide guidance on how many variables to use in the prediction model. In this case, it is possible to use as few as two and as many as 15 variables in the models. The largest difference between the correlations is between the first and second variables, and after that the differences are smaller. Thus, if a low number of variables is preferred, it is possible to use two variables. The second sharp drop is between the 15th and 16th variables. Thus if a higher number of variables is desired, it is possible to choose 15 variables. After the 15th variable, the correlation drops below 0.3. A Pearson correlation coefficient of less than 0.3 is not a high correlation (Field, 2009). Using a variable with a correlation below 0.3 may not increase the accuracy of the predictions and it is not recommended for use in the model. However, adding all of these 15 variables will not necessary increase the accuracy of the prediction and a subset of these variable may results in greater predictive accuracy, especially for predicting at-risk students in future semesters. In the next chapter, the accuracy of predictions using different number of variables will be tested.
3.6.2 Gini gain

The second feature selection method that was used in this study was the Gini gain method. The goal of this method is to calculate how well a variable can divide the students into pass or fail categories. This method is based on maximizing the Gini score. The higher a variable’s Gini score, the better it divides the students into pass or fail categories. A Gini score is calculated using a Gini function before and after dividing the dataset into two or more categories (e.g., pass and fail students) based on a predictor variable. The Gini function is used to calculate the purity of data in each category (Breiman et al., 1984). For example, a variable that partitions the students into two categories, in which all passed students are in one partition and all failed students are in another partition, will result in the highest possible Gini score. For each variable a Gini score was calculated. Variables with the highest Gini score, meaning better separate passed from failed students, were selected. For a given dataset $S$ and a variable $A$, the Gini gain is calculated as:
\[
Gain(S, A) = Gini(S) - \sum_{v \in \text{values}(A)} \frac{|S_v|}{|S|} Gini(S_v) \quad (Eq. 11)
\]

\[
Gini(S) = 1 - \sum_{i=1}^{k} p_i^2 \quad (Eq. 12)
\]

where:

- \( S \) is the dataset containing all variables,
- \( A \) is a variable in the \( S \) dataset,
- \( v \) is a possible value of variable \( A \),
- \( S_v \) is the a subset of \( S \) where \( A=v \),
- \( p_i \) is the probability of being in category \( i \), and
- \( k \) is the number of categories.

Figure 3.3 shows the Gini gain of the 44 predicting variables. The aim is to select the variables with higher Gini gain. Based on the scree plot, it is possible to choose two, four, seven, or 13 variables for use in the prediction models. In the next chapter, the accuracy of predictions using different numbers of variables will be tested to show how choosing different numbers of variables affects accuracy of the predictions.
Figure 3.3 – Gini gain of the predicted variables.

3.6.3 Sum of Squared Errors (SSE) and explained variance (R²)

The third feature selection method used in this study was based on calculating the Sum of Squared Errors (SSE) and the percentage of variance of the predicted variable (pass/fail) explained by the predicting variables. Variables that result in lower SSE and higher explained variance are better predictors of at-risk students in the course. A series of Generalized Linear Models (GLMs) were created by sequentially selecting the predictor variables. For each model, SSE and explained variance, $R^2$, were calculated (Eqs. 13-15).

$$SSE = \sum_{i} (y_i - \hat{y}_i)^2 \quad \text{(Eq. 13)}$$

$$SST = \sum_{i} (y_i - \bar{y})^2 \quad \text{(Eq. 14)}$$

$$R^2 = \frac{SSE}{SST} \quad \text{(Eq. 15)}$$

where
\( y_i \) is the actual value of the predicted variable (e.g., final grade of a student),
\( \hat{y}_i \) is the estimated value of the predicted variable by GLM model, and
\( \bar{y} \) is the mean of the actual values of the predicted variable.

Using more variables in the prediction model will result in lower SSE and higher explained variance. However, increasing the number of predictor variables increases the complexity of the model. This increases the time taken to train the models, and more importantly may decrease the accuracy of the predictions for future data, because some of the variables may not be good predictors of success.

Figure 3.4 shows the explained variance \((R^2)\) after adding each variable to the model. As it can be seen, the first two variables result in the maximum increase in the explained variance. The increase in the explained variance after the first five variables is very low. Thus choosing two or five variables may result in high accuracy predictions. Choosing more than five variables will only increase the complexity of the model without significantly improving its accuracy. Typically complex models do not have high accuracy for future unseen data, because they fit too closely to the training data. The accuracy of different models with different number of variables will be tested in Chapter 4.
3.6.4 Optimal number of variables

As it can be seen from the previous sections, different feature selection methods may result in different numbers of variables. In addition, even with using one feature selection method, it is possible to choose different numbers of variables. As explained earlier, there is a trade off between the complexity of the models and the number of variables. A lower number of variables typically results in more generalizable models with the potential for greater accuracy when applied to future data. Thus my suggestion is to select the lowest number of variables, which is two. In the next chapter, the top two prediction models with different number of variables are trained, verified, and tested to showcase the differences in the accuracy of these models.

3.7 Model robustness

Every modeling method has a training size range in which the models created by the method works accurately. Some models are more sensitive to the
training size than others. A training size that is too small or too large may decrease the accuracy of the predictions. A too small training size may not have enough information about the relationship between the variables to create the models. A too large sample size may result in building a model that is too close to the training data and is not a good fit for future unseen data. To identify the low and high training size boundaries of the models developed in this study, different subsets of the train dataset were used to train the models. Then the models’ accuracies were evaluated. This analysis specified the dataset range (i.e., number of students in the course) that the models can be trained with. If the class size is smaller or bigger than these limits, the models developed in this study may not be as accurate for predicting at-risk students.

To create training sets with different sizes, the training dataset with 780 students was randomly divided into 20 clusters of 39 students (keeping the pass/fail ratio the same). A smaller cluster size leads to a more accurate estimate of the training size range for the modeling methods. The smallest possible cluster size in this study was 39. Because fewer than 10% of students failed the course, a cluster with 39 students has only 3 or 4 students who failed the course. This was the lowest number of students in each category that was required for some of the prediction modeling methods to train the models.

To change the training size for the models, the number of the clusters was increased from one to 20 in the training phase. This resulted in training the models with 39 to 780 students.
3.8 Workflow of model development

Figure 3.5 shows the overall workflow of the model development. Three main phases that are color coded in the diagram and refer to: (1) selecting the top two prediction modeling methods, (2) selecting the best variables for identifying at-risk students, and (3) identifying the range for which the models are working accurately by evaluating the models’ robustness to training (i.e., class) size.

In the first phase, the six modeling methods introduced in section 3.4 were used to train and verify the prediction models. After error analysis of these six created models, an ensemble modeling method was used to create a model with reduced error. Then the top two prediction modeling methods were selected. In addition, the predictive power of different assessments in the course, including homework learning objective scores, quiz grades, and midterm exam grade, were investigated. In the second phase, the three feature selection methods, introduced in section 3.6, were used to find the best variables for identifying at-risk students. The top two modeling methods, identified in phase 1, were used to train models with different numbers of variables. Then these models were verified and tested to select the optimal number of variables. In the third phase, the training data were divided into 20 clusters and used along with the top two modeling methods (from phase 1) to train models with the optimal number of variables (from phase 2) to find the training size range (i.e., class size) in which the models are perform accurately.
Figure 3.5 - Overall workflow of model development.
CHAPTER 4 - MODEL DEVELOPMENT

Using the data from Spring 2013 and 2014, I trained, verified, and tested the models created by the predictive modeling methods introduced in Chapter 3. The methods were Logistic Regression (Log Reg), K-Nearest Neighbor (KNN), Multi-Layer Perceptron (MLP) Neural Network, Decision Tree (DT), Support Vector Machine (SVM), and Naive Bayes Classifier (NBC). Three feature selection methods, correlations, Gini gain, and explained variance, were used to select the variables with the most predictive power.

4.1 Training and verifying the models

The six predictive modeling methods introduced in Chapter 3 were used to train prediction models. The training dataset were from Spring 2013 containing 780 students. Then the models were utilized to identify at-risk students for the Verify 1 dataset, containing 390 students. The results of testing the six different methods with all 44 variables are reported in Table 4.1. Logistic Regression, which is the most popular prediction modeling method in educational settings, was used as the baseline method. After training with the train dataset, Logistic Regression was utilized to predict pass/fail students for the Verify 1 dataset. Logistic Regression’s $F_{1.5}$ score was 0.56 with the overall accuracy of 92.6%. Its accuracy for students who
passed the course is 95.3%, and for students who failed the course is 58.6%. Because the course grade distribution is negatively skewed, this method performs significantly better for students who passed the course.

Table 4.2 - Verification results and accuracy of prediction models.

<table>
<thead>
<tr>
<th>Characteristics</th>
<th>Methods</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Log Reg</td>
</tr>
<tr>
<td>$F_{1.5}$</td>
<td>0.56</td>
</tr>
<tr>
<td>Accuracy</td>
<td>92.6%</td>
</tr>
<tr>
<td>Accuracy-Pass</td>
<td>95.3%</td>
</tr>
<tr>
<td>Accuracy-Fail</td>
<td>58.6%</td>
</tr>
<tr>
<td>True Negative</td>
<td>344</td>
</tr>
<tr>
<td>False Positive</td>
<td>17</td>
</tr>
<tr>
<td>False Negative</td>
<td>12</td>
</tr>
<tr>
<td>True Positive</td>
<td>17</td>
</tr>
</tbody>
</table>

4.1.1 Misidentifications and errors

To better understand the misidentifications for each model, I examined the distribution of misidentifications based on the actual end-of-semester letter grades (Figure 4.1). Two types of misidentifications are false positives and false negatives. False positives (i.e., type I errors) are students who pass the course but identified as at-risk students by the models. False negatives (i.e., type II errors) are students who fail the course but are not identified as at-risk students by the models.

While it is not possible to have zero false positive or negative errors (i.e., no type I or type II errors), a model that does not identify, for example, an F grade student as a potential pass is preferred to the one that does. Thus, it is more important to minimize false negative or type II errors. A good model is a model for which most of the error happens in predicting passes and fails around the C and D divide. However, from an educational perspective, it is better to misidentify a C
student (i.e., have a false positive or type I error) than a D student (i.e., have a false negative or type II error). Misidentifying a C student as a potential fail may encourage the student to work harder and improve his/her grade to a B. Misidentifying a D student as a potential pass may prevent him/her from taking actions to pass the course.

![Graphs showing misidentifications by letter grade](image)

Figure 4.1 - Number of misidentifications based on the letter grades.

### 4.1.2 Best models for predicting students who passed

Based on the analyses, the best model for overall accuracy was the model created by the K-Nearest Neighbor (KNN) method, which identified 94.9% of the students (failed and passed combined) correctly (Table 4.1). This model was also the best for predicting the students who passed the course. The only false positive
misidentification by the KNN model was a C grade. Thus, it had 99.7% accuracy for students who passed the course. However, the KNN model performed poorly at identifying at-risk students correctly and had the lowest $F_{1.5}$ score. The KNN model identified only 10 out of 29 at-risk students and had 34.5% accuracy for failure students, making it the worst model for identifying at-risk students in this study. The KNN model had the highest number of misidentifications of D and F grades (Figure 4.1). The second best model overall and for identifying students who passed the course was the model created by Multi-Layer Perceptron Neural Network (MLP) method, with overall accuracy of 93.1%, which was close to the KNN model. The MLP model accuracy at identifying passed students was 96.7%, which was 3% lower than the KNN model. This model's accuracy at identifying students who failed the course was 48.3%, which is higher than the KNN model. The model created by the Decision Tree (DT) method, was the third best model for identifying students who passed the course correctly, which was very similar to the MLP model. However, the DT model accuracy at predicting at-risk students (44.8%) was lower than the MLP model.

4.1.3 Best models for predicting students who failed

Based on the analysis, the best model for identifying at-risk students correctly was the model created by the Naive Bayes Classifier (NBC) method, which identified 86.2% of students who failed the course. The NBC model only misidentified four at-risk students and had the highest $F_{1.5}$ score. However, the NBC model had the lowest overall accuracy (86.9%) and lowest accuracy for identifying
students who passed the course (87.0%). The second best model for identifying at-risk students was the model created by the Support Vector Machine (SVM) method. This model’s overall accuracy and accuracy for students who passed the course were similar to the NBC model. However, the SVM model’s accuracy at identifying at-risk students was less than that for the NBC model; the SVM model could not identify eight out of 29 at-risk students. The NBC model and then the SVM model had the lowest misidentification of D and F grades. While both of these models had high false positive error, the SVM model performed better for A and C students, and the NBC model performed better for B students.

### 4.2 Creating an Ensemble model

As can be seen in Table 4.1 and Figure 4.1, none of the prediction models had acceptable accuracy for both students who passed and who failed the course. One possible way to improve the predictions was to create an Ensemble model using three of the models. To make an Ensemble model, first I took a closer look at the students’ grade distribution and number of times a student was misidentified for all six models (Figure 4.2). For example, a student (denoted with a circle in Figure 4.2) with an x-axis value of 54 and a y-axis value of six had a final grade of 54 (fail) and was misidentified by all six models; this is a false negative error. Three of the misidentified students were not identified by any of the six models. Six students were only identified by one model. The rest of the students were identified by at least two of the six models. Thus, it might be possible to improve the predictive power by using three of the models as part of an Ensemble model. To keep the false
negative errors low, which is important for the classifications, but at the same time to decrease the false positive errors, I used the two models with the lowest false negative errors, the NBC and SVM models, and the model with lowest false positive errors, the KNN model.

Figure 4.2 - Number of times a student has been misidentified by the six models.

The results of testing the Ensemble model are reported in Table 4.2 and Figure 4.3. The $F_{1.5}$ score for the Ensemble model was 0.61, which was higher than all other models. The accuracy of the Ensemble model for students who passed the course was 93.9%, which was lower than the KNN model but higher than the SVM and NBC models. The accuracy of the Ensemble model for students who failed the course was 69%, which was two times more than the KNN model (34.5%) yet lower than the NBC model (86.2%) and close to the SVM model (72.4%). The false negative misidentifications of Ensemble model (9 students) were similar to the SVM model (8 students), which was the second best method for identifying at-risk
students. The Ensemble model's false positive misidentifications (22 students) was 52% of the SVM model (42 students). In addition, most of the misidentifications were C grade students, which is the best possible outcome among the possible errors. However, this model misidentified one A and three F students.

<table>
<thead>
<tr>
<th>Method</th>
<th>Ensemble</th>
</tr>
</thead>
<tbody>
<tr>
<td>$F_{1.5}$</td>
<td>0.61</td>
</tr>
<tr>
<td>Accuracy</td>
<td>92.1%</td>
</tr>
<tr>
<td>Accuracy-Pass</td>
<td>93.9%</td>
</tr>
<tr>
<td>Accuracy-Fail</td>
<td>69.0%</td>
</tr>
<tr>
<td>True Negative</td>
<td>339</td>
</tr>
<tr>
<td>False Positive</td>
<td>22</td>
</tr>
<tr>
<td>False Negative</td>
<td>9</td>
</tr>
<tr>
<td>True Positive</td>
<td>20</td>
</tr>
</tbody>
</table>

4.3 Predictive power of different assessments

To investigate the extent to which each assessment contributes to the prediction of at-risk students, the models using subsets of the training datasets were created. Since some courses do not have a midterm exam until after the middle of the semester, it is important to know how much just quiz grades and learning objective scores, which are collected weekly, contribute to identifying at-risk students. Midterm exam grades were removed from the dataset and the prediction models were created and tested (Table 4.3). As was expected given the weight of the midterm exam in the overall course grade, the performance of the models decreased from the best $F_{1.5}$ of 0.61 to 0.57. The NBC and Ensemble models were still the best
prediction models. Removing the midterm exam grade and only using homework learning objective scores and quiz grades did not significantly decrease the accuracy of the models. Therefore, it is possible to identify at-risk students with good accuracy without the first midterm exam grade and only based on learning objective scores and quiz grades.

Table 4.3 – $F_{1.5}$ scores of the models created with different assessment grades.

<table>
<thead>
<tr>
<th>Assessments</th>
<th># of variables</th>
<th>Log Reg</th>
<th>DT</th>
<th>SVM</th>
<th>MLP</th>
<th>KNN</th>
<th>NBC</th>
<th>Ensemble</th>
</tr>
</thead>
<tbody>
<tr>
<td>All</td>
<td>44</td>
<td>0.56</td>
<td>0.43</td>
<td>0.50</td>
<td>0.46</td>
<td>0.53</td>
<td>0.59</td>
<td>0.61</td>
</tr>
<tr>
<td>Quiz &amp; HW</td>
<td>43</td>
<td>0.52</td>
<td>0.50</td>
<td>0.52</td>
<td>0.40</td>
<td>0.46</td>
<td>0.56</td>
<td>0.57</td>
</tr>
<tr>
<td>HW</td>
<td>33</td>
<td>0.35</td>
<td>0.41</td>
<td>0.45</td>
<td>0.27</td>
<td>0.43</td>
<td>0.52</td>
<td>0.50</td>
</tr>
<tr>
<td>Quiz</td>
<td>10</td>
<td>0.35</td>
<td>0.37</td>
<td>0.44</td>
<td>0.47</td>
<td>0.31</td>
<td>0.39</td>
<td>0.39</td>
</tr>
<tr>
<td>Midterm</td>
<td>1</td>
<td>0.49</td>
<td>0.51</td>
<td>0.49</td>
<td>0.42</td>
<td>0.35</td>
<td>0.60</td>
<td>0.60</td>
</tr>
</tbody>
</table>

The models using only learning objective scores were created to demonstrate learning objectives’ ability to identify at-risk students. Removing quiz grades from the models resulted in a decrease in the $F_{1.5}$ scores (0.52 for the best model). The NBC and Ensemble models had the highest $F_{1.5}$ scores. To compare the quiz grade and learning objective scores, the models using only quiz grades were created. Only using quiz grades decreased the models performance considerably, the best model has an $F_{1.5}$ score of 0.47. Compared to quiz grades, learning objective scores were a better predictor of students’ success or failure in the course.

Finally, models were created using only the midterm exam grade to identify at-risk students. Using only the midterm exam grade, the accuracy of the models was similar to the models using all variables. The NBC and Ensemble were the best prediction models in this case.
Because midterm exam grades were the best predictor of at-risk students, I also predicted students’ success on the midterm exam using the first three weeks learning objectives scores and quiz grades. Overall, the accuracy of predictions for students who failed the midterm exam was low (Table 4.4). The SVM followed by the NBC were the best two models in predicting which students failed the midterm exam. The KKN model was the best model for predicting students who were successful on midterm exam.

Table 4.4 – Accuracy of predicting success at midterm exam.

<table>
<thead>
<tr>
<th>Characteristics</th>
<th>Methods</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Log Reg</td>
</tr>
<tr>
<td>$F_{1.5}$</td>
<td>0.37</td>
</tr>
<tr>
<td>Accuracy</td>
<td>89.5%</td>
</tr>
<tr>
<td>Accuracy-Pass</td>
<td>97.7%</td>
</tr>
<tr>
<td>Accuracy-Fail</td>
<td>31.3%</td>
</tr>
<tr>
<td>True Negative</td>
<td>334</td>
</tr>
<tr>
<td>False Positive</td>
<td>8</td>
</tr>
<tr>
<td>False Negative</td>
<td>33</td>
</tr>
<tr>
<td>True Positive</td>
<td>15</td>
</tr>
</tbody>
</table>

4.4 Feature selection

Three feature selection methods, correlations, Gini gain, and explained variance (described in section 3.6 in more details), were used in this study. Because the models created by the NBC and Ensemble methods consistently performed better than other models, and the focus of this section is on selecting the variables, only these two modeling methods were used. The models were built based on different numbers of variables using the training dataset, consisting of 780 students from Spring 2013, and verified using verify 2 dataset, consisting of 390 students.
from the same semester. To create the models, variables entered with three different sequences as described below.

First, using the correlation method, the variables were sorted based on their correlation with the outcome variable (i.e., pass or fail the course). The models were created by adding one predictor variable at a time. The variables with highest correlations were added to the model first. Figure 4.4 shows the verification results of these models. As it can be seen, the best $F_{1.5}$ score was achieved with only three variables in the model including midterm exam grade and two learning objectives.

![Figure 4.4](image-url)  
**Figure 4.4** – Accuracy of the models using different numbers of variables. Variables were sorted and added to the model based on correlation with passing/failing the course.

Second, using the Gini gain method, the variables were sorted based on their Gini gain score. The variables with the highest gains were added to the model first. Figure 4.5 shows the verification results for these models. Similar to the correlation method, the best $F_{1.5}$ score was achieved with only three variables in the model, midterm exam grade and two learning objectives. These two learning objectives
were similar to the top two learning objectives that were used in the correlation feature selection method. These learning objectives are discussed in Chapter 5.

Figure 4.5 – Accuracy of the models using different number of variables. Variables were sorted and added to the model based on their Gini gain.

Third, using the explained variance method, the variables were sorted based on the amount of explained variance in the outcome variables (i.e., pass or fail the course). The variables with highest explained variances were added to the model first. Figure 4.6 shows the verification results of these models. As it can be seen, the best $F_{1.5}$ was score achieved with only four variables in the model, the midterm exam grade and three learning objectives.
Figure 4.6 – Accuracy of the models using different number of variables. Variables were sorted and added to the model based on the variance they explained in the outcome variable.

As it can be seen in Figures 4.4 to 4.6, using more variables did not improve the accuracy of the predictions. The NBC and Ensemble models' accuracy was very similar (the $F_{1.5}$ difference was less than 0.02) for the first five variables. The best $F_{1.5}$ score for all methods was close to 0.70. The top three variables for the correlation and Gini gain methods, which resulted in the highest accuracy in the predictions, were the same. Thus using these three variables with the NBC or Ensemble models can result in the highest accuracy in the predictions.

4.5 Model robustness

After identifying the best modeling methods and the optimal number of variables to use in them, the robustness of the selected modeling methods to the size of the training dataset was investigated. Limited data size may result in low accuracy of the models. However, very large training dataset may result in
overfitting the models to the training dataset and also decrease the accuracy of the predictions.

The models were trained with the top three variables selected from the correlation feature selection method (the midterm exam and two learning objective grades). After the dataset size passed about 120 students, increasing the number of students in the training dataset did not improve the accuracy of the results (Figure 4.7). Increasing the size of the training dataset to 780 yielded similar results to 120.

![Figure 4.7 – Accuracy of the models based on different training size.](image)

### 4.6 Testing the models performance

The performance of the top two prediction models (NBC and Ensemble) at identifying at-risk students in the next semester’s data was tested. Similar to the previous sections, the models were trained based on the train dataset, which consisted of 780 students from Spring 2013. The test dataset consisted of 1413 students from Spring 2014. The accuracy of the models on this test dataset shows how effective these models are in identifying at-risks students in future semesters.
The course syllabus and grading scheme in Spring 2013 and Spring 2014 were very similar. There were some minor changes from one semester to the other. Some of the instructional team members including instructors and teaching assistants were changed from Spring 2013 to Spring 2014. However, the majority of the course syllabus elements were the same between the two semesters. If there were major changes in the course, these models may not be able to identify at-risk students in the subsequent semester with high accuracy.

Similar to the previous section, to create the models, variables were sorted and used in the models based on the three feature selection methods: (1) correlations, (2) Gini gain, and (3) explained variance. The models were then tested with the Spring 2014 test dataset. Figures 4.8, 4.9, and 4.10 show the results of the tests for correlation, Gini gain, and explained variance feature selection methods. These results show whether or not the optimal number of variables is the same for the verify dataset (from same semester) and the test dataset (from future semester).

For all feature selection methods, the top models’ (NBC and Ensemble) $F_{1.5}$ scores were the same for the first six variables. For all three methods, the $F_{1.5}$ score decreased after two or three variables and then gradually increased. Among the three feature selection methods, correlation achieved the highest $F_{1.5}$ score with only two variables, which was close to 0.7. Achieving a similar $F_{1.5}$ score between the same semester dataset and the next semester test dataset indicates that the models’ performance were similar in identifying at-risk students for the same semester data and for the future semester data. Thus, it is possible to train the models with one
semester’s worth of data and then use the models in the next semester to identify at-risk students.

Figure 4.8 – Accuracy of the models for future data using different numbers of variables. Variables were sorted and added to the model based on their correlation with pass/fail the course.

Figure 4.9 – Accuracy of the models for future data using different numbers of variables. Variables were sorted and added to the model based on Gini gain.
Figure 4.10 – Accuracy of the models for future data using different numbers of variables. Variables were sorted and added to the model based on the variance they explained in the outcome variable.

Models with fewer variables are preferred to the models with a greater number of variables. Models consisting of a high number of variables are more complex, need more time to train, and are typically less generalizable. As it can be seen in Figures 4.8 to 4.10, adding variables after the second or third variables does not increase the performance of the models significantly. Thus two or three is the optimal number of variables to use in the models.

Table 4.5 shows the accuracy results for the top two models with two variables selected by the correlation method, which had the best results. The NBC and Ensemble models performed similarly at identifying at-risk students in the next semester. The overall accuracy of these models were 90.1%. With an $F_{1.5}$ score of 0.67, these models identified 77.6% of at-risk students correctly.
Table 4.5 - Test results of top two models using two variables selected by the correlation method.

<table>
<thead>
<tr>
<th>Characteristics</th>
<th>NBC</th>
<th>Ensemble</th>
</tr>
</thead>
<tbody>
<tr>
<td>$F_{1.5}$</td>
<td>0.67</td>
<td>0.67</td>
</tr>
<tr>
<td>Accuracy</td>
<td>90.1%</td>
<td>90.1%</td>
</tr>
<tr>
<td>Accuracy-Pass</td>
<td>91.5%</td>
<td>91.5%</td>
</tr>
<tr>
<td>Accuracy-Fail</td>
<td>77.6%</td>
<td>77.6%</td>
</tr>
<tr>
<td>True Negative</td>
<td>1159</td>
<td>1159</td>
</tr>
<tr>
<td>False Positive</td>
<td>107</td>
<td>107</td>
</tr>
<tr>
<td>False Negative</td>
<td>33</td>
<td>33</td>
</tr>
<tr>
<td>True Positive</td>
<td>114</td>
<td>114</td>
</tr>
</tbody>
</table>

Figure 4.11 shows the distribution of misidentifications for different letter grades for the top two models. Both models performed the same. Most of the misidentifications (52%) were for C grades, which is the best possible grade for misidentifications. As it was discussed earlier, while it is not possible to reduced the misidentifications to zero, a good model is a model for which most of the error happens when predicting passes and fails around the C and D divide. However, from an educational perspective, it is better to misidentify a C student (i.e., have a false positive or type I error) than a D student (i.e., have a false negative or type II error). Misidentifying a C student as a potential fail may encourage the student to work harder and improve his/her grade to a B. Although these models could not identify 33 students who were at-risk, only 11 of these students actually failed the course with a F letter grade or withdraw from the course.
Figure 4.11 - Number of misidentifications for the top two models.
CHAPTER 5 - LEARNING OBJECTIVES

In this chapter, I review the course learning objectives and the correlation among them and other course assessments. I also review learning objectives that are related to students’ success in the course and possible threshold learning objectives for the course.

5.1 Learning objectives used in the models

Table 5.1 shows the list of homework learning objectives for weeks 1-5 that were used in the models. Each homework had six or seven learning objectives. All homeworks had "coding standards" and "professional habits" learning objectives. The coding standards learning objective referred to how well students followed the course established standards for coding including proper commenting and variable naming. The professional habits learning objective refers to how well students followed overall homework instructions including proper use of the course template, use of specified file names, and completing of all the necessary submission components.

In this course, homework 1 learning objectives were a review of a prerequisite course’s learning objectives. Also, none of the learning objectives in the course were built on homeworks 3 and 4 learning objectives. Thus in the first five
weeks of the semester, only homeworks 2 and 5 learning objectives were constantly being used in the next weeks.

Table 5.1 – Description of homework learning objectives for weeks 1 to 5 that were used in the prediction models.

<table>
<thead>
<tr>
<th>Homework</th>
<th>LO</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>HW01</td>
<td>1</td>
<td>Perform complex calculations using algebraic and trigonometric functions in computations with scalars vectors and matrices.</td>
</tr>
<tr>
<td>HW01</td>
<td>2</td>
<td>Identify appropriate uses for dot notation.</td>
</tr>
<tr>
<td>HW01</td>
<td>3</td>
<td>Demonstrate use of colon operator for creating data structures.</td>
</tr>
<tr>
<td>HW01</td>
<td>4</td>
<td>Demonstrate how to create a MATLAB plot for technical presentation.</td>
</tr>
<tr>
<td>HW01</td>
<td>5</td>
<td>Demonstrate how to plot multiple data sets on one figure.</td>
</tr>
<tr>
<td>HW01</td>
<td>6</td>
<td>Coding Standards</td>
</tr>
<tr>
<td>HW01</td>
<td>7</td>
<td>Professional Habits</td>
</tr>
<tr>
<td>HW02</td>
<td>1</td>
<td>Use basic relational operators.</td>
</tr>
<tr>
<td>HW02</td>
<td>2</td>
<td>Demonstrate the ability to create a histogram for technical presentation.</td>
</tr>
<tr>
<td>HW02</td>
<td>3</td>
<td>Interpret and evaluate logical statements.</td>
</tr>
<tr>
<td>HW02</td>
<td>4</td>
<td>Construct logical statements from English statements.</td>
</tr>
<tr>
<td>HW02</td>
<td>5</td>
<td>Coding Standards</td>
</tr>
<tr>
<td>HW02</td>
<td>6</td>
<td>Professional Habits</td>
</tr>
<tr>
<td>HW03</td>
<td>1</td>
<td>Demonstrate the ability to create a histogram for technical presentation using Excel.</td>
</tr>
<tr>
<td>HW03</td>
<td>2</td>
<td>Construct cumulative distribution plots in Excel.</td>
</tr>
<tr>
<td>HW03</td>
<td>3</td>
<td>Interpret cumulative distribution plots.</td>
</tr>
<tr>
<td>HW03</td>
<td>4</td>
<td>Construct and interpret cumulative distribution plots in MATLAB.</td>
</tr>
<tr>
<td>HW03</td>
<td>5</td>
<td>Construct a figure window to display multiple plots and/or histograms.</td>
</tr>
<tr>
<td>HW03</td>
<td>6</td>
<td>Coding Standards</td>
</tr>
<tr>
<td>HW03</td>
<td>7</td>
<td>Professional Habits</td>
</tr>
<tr>
<td>HW04</td>
<td>1</td>
<td>Interpret cumulative distribution plots (Problem 1).</td>
</tr>
<tr>
<td>HW04</td>
<td>2</td>
<td>Generate a histogram suitable for technical presentation from a cumulative distribution plot.</td>
</tr>
<tr>
<td>HW04</td>
<td>3</td>
<td>Construct cumulative distribution plots in Excel.</td>
</tr>
<tr>
<td>HW04</td>
<td>4</td>
<td>Interpret cumulative distribution plots (Problem 2).</td>
</tr>
<tr>
<td>HW04</td>
<td>5</td>
<td>Construct cumulative distribution plots in MATLAB.</td>
</tr>
<tr>
<td>HW04</td>
<td>6</td>
<td>Coding Standards.</td>
</tr>
<tr>
<td>HW04</td>
<td>7</td>
<td>Professional Habits.</td>
</tr>
<tr>
<td>HW05</td>
<td>1</td>
<td>Construct an appropriate function definition.</td>
</tr>
<tr>
<td>HW05</td>
<td>2</td>
<td>Demonstrate how to create a MATLAB plot for technical presentation.</td>
</tr>
<tr>
<td>HW05</td>
<td>3</td>
<td>Create a user-defined function using rules for writing functions.</td>
</tr>
<tr>
<td>HW05</td>
<td>4</td>
<td>Create test cases to evaluate a user-defined function.</td>
</tr>
<tr>
<td>HW05</td>
<td>5</td>
<td>Coding Standards.</td>
</tr>
<tr>
<td>HW05</td>
<td>6</td>
<td>Professional Habits.</td>
</tr>
</tbody>
</table>
Figure 5.1 shows the distribution and average learning objective scores that were used in the prediction models for students who passed and failed the course. For all learning objectives, the differences between the average score of failing and passing students were statistically significant (t-test p-value < 0.05). For successful students, most of the average scores were more than 2 out of 3. In addition, at least half of the students who passed the course received the full score of 3 on 28 out of the 33 learning objectives. Only five learning objectives, three from homework 4 and two from homework 3, had a median of 2. For at-risk students, most of the average learning objective scores were less than 2. Most at risk students did well on homeworks 1 and 2; at least half of at-risk students learning scores were 2 or better. At-risk students also did well on most of homework 3 learning objectives.

Figure 5.1 – Distribution and average learning objective scores for weeks 1-5 for students who passed and failed (* illustrates leaning objective mean).
Successful and at-risk students distribution of grades in homework 5 learning objectives were different. While the majority of successful students received a 3 score on homework 5 learning objectives, the majority of at-risk students received a score of zero. In week 5, students learned user-defined functions in MATLAB.

Figure 5.2 shows the correlation between academic success in the course and the course assessments including midterm exam, homework learning objectives, and quizzes. Blue indicates a positive correlation, a darker blue shows a higher correlation, and red shows a negative correlation. Because the midterm exam grade was selected by all three feature selection methods, it was expected that the midterm exam and being successful in the course would be highly correlated. The correlation between the homework learning objectives and course success increased from homework 1 to 5. Homework 1 learning objectives had very low correlation (less than 0.20) with success in the course. Homework 2 learning objectives also had low correlation (less than 0.30) with success in the course. The correlations of homework 3 learning objectives with course success were low or moderate (less than 0.32). Homework 4 learning objectives were split into two categories. The first four learning objectives had low correlation with success in the course (less than 0.26) while the last 3 had moderate correlations. Homework 5 learning objectives had moderate or good correlations (they ranged from 0.31 to 0.42) with the course success.
Figure 5.2 – Correlation among course assessments that were used in the prediction models (first five weeks of the semester). Numbers rounded to the nearest tenth. A cell with only a dot and no digits shows a correlation rounded to 0 (image best viewed in color).
All the learning objectives in a homework were correlated to each other (typically more than 0.3), with the exception of homework 4. Learning objectives HW04-1 and HW04-4 (Interpret cumulative distribution plots for problem 1 and 2) not only were not correlated with other homework 4 learning objectives but they also had negative correlation with homework 3 learning objectives (as high as -0.29). In addition, learning objective HW01-1 (Perform complex calculations using algebraic and trigonometric functions in computations with scalars vectors and matrices) had very low negative correlation with three learning objectives, HW02-4 (Construct logical statements from English statements), HW03-1 (Demonstrate the ability to create a histogram for technical presentation using Excel), and HW03-3 (Interpret cumulative distribution plots).

Quizzes did not follow a pattern in their correlation with the course success. Only three quizzes (2.a, 3.a, and 4.a) had correlation higher than 0.3 with the course success. Quiz 1.a, 1.b, and 5.b had negative correlations with homework 3 learning objectives. In addition, all quizzes had very low or negative correlations with learning objective HW01-1.

5.2 Learning objectives after week 5

Table 5.2 lists the homework learning objectives after week 5. Figure 5.3 shows the correlations among homework leaning objectives for the entire semester. Similar to the first five weeks, learning objectives on any given homework were correlated to each other. Only one learning objective, HW13-4 (Use Excel to plot a linear trend line over the raw data and display the equation and r-squared value),
<table>
<thead>
<tr>
<th>Homework</th>
<th>LO</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>HW06</td>
<td>1</td>
<td>Draw and interpret flow charts containing decision branches to characterize an engineering problem.</td>
</tr>
<tr>
<td>HW06</td>
<td>2</td>
<td>Write evidence based rationales.</td>
</tr>
<tr>
<td>HW06</td>
<td>3</td>
<td>Demonstrate the proper method to pass information to/from a function.</td>
</tr>
<tr>
<td>HW06</td>
<td>4</td>
<td>Coding Standards.</td>
</tr>
<tr>
<td>HW06</td>
<td>5</td>
<td>Professional Habits.</td>
</tr>
<tr>
<td>HW07</td>
<td>1</td>
<td>Given a verbal description of the logic underlying a conditional decision write the flowchart illustrating the logical flow.</td>
</tr>
<tr>
<td>HW07</td>
<td>2</td>
<td>Given a flowchart containing conditional statements create MATLAB code to implement the logic.</td>
</tr>
<tr>
<td>HW07</td>
<td>3</td>
<td>Describe regions of a graph using equalities and inequalities.</td>
</tr>
<tr>
<td>HW07</td>
<td>4</td>
<td>Develop appropriate logic conditions and conditional statements based on a problem statement.</td>
</tr>
<tr>
<td>HW07</td>
<td>5</td>
<td>Create test cases to evaluate a user-defined function.</td>
</tr>
<tr>
<td>HW07</td>
<td>6</td>
<td>Coding Standards.</td>
</tr>
<tr>
<td>HW07</td>
<td>7</td>
<td>Professional Habits.</td>
</tr>
<tr>
<td>HW08</td>
<td>1</td>
<td>Create a while loop flowchart from a problem description.</td>
</tr>
<tr>
<td>HW08</td>
<td>2</td>
<td>Write MATLAB while loop code from a flowchart.</td>
</tr>
<tr>
<td>HW08</td>
<td>3</td>
<td>Utilize variable tracking tables to predict values of each variable throughout the execution of a loop.</td>
</tr>
<tr>
<td>HW08</td>
<td>4</td>
<td>Create a for loop flowchart from a problem description.</td>
</tr>
<tr>
<td>HW08</td>
<td>5</td>
<td>Write MATLAB for loop code from a flowchart.</td>
</tr>
<tr>
<td>HW08</td>
<td>6</td>
<td>Coding Standards.</td>
</tr>
<tr>
<td>HW08</td>
<td>7</td>
<td>Professional Habits.</td>
</tr>
<tr>
<td>HW09</td>
<td>1</td>
<td>Create a nested loop flowchart from a problem description.</td>
</tr>
<tr>
<td>HW09</td>
<td>2</td>
<td>Write MATLAB nested loop code from a flowchart.</td>
</tr>
<tr>
<td>HW09</td>
<td>3</td>
<td>Create appropriate test case for a given algorithm.</td>
</tr>
<tr>
<td>HW09</td>
<td>4</td>
<td>Create a while loop flowchart from a problem description.</td>
</tr>
<tr>
<td>HW09</td>
<td>5</td>
<td>Write MATLAB while loop code from a flowchart.</td>
</tr>
<tr>
<td>HW09</td>
<td>6</td>
<td>Coding Standards.</td>
</tr>
<tr>
<td>HW09</td>
<td>7</td>
<td>Professional Habits.</td>
</tr>
<tr>
<td>HW11</td>
<td>1</td>
<td>Create a GUI flowchart from a problem description.</td>
</tr>
<tr>
<td>HW11</td>
<td>2</td>
<td>Construct a GUI layout (Problem 1).</td>
</tr>
<tr>
<td>HW11</td>
<td>3</td>
<td>Use Property Inspector to manually change the Tag property for a GUI component.</td>
</tr>
<tr>
<td>HW11</td>
<td>4</td>
<td>Construct a GUI layout (Problem 2).</td>
</tr>
<tr>
<td>HW11</td>
<td>5</td>
<td>Write code to enable each component of the GUI layout to function appropriately.</td>
</tr>
<tr>
<td>HW11</td>
<td>6</td>
<td>Coding Standards.</td>
</tr>
<tr>
<td>HW11</td>
<td>7</td>
<td>Professional Habits.</td>
</tr>
<tr>
<td>HW12</td>
<td>1</td>
<td>Create a GUI flowchart from a problem description (Problem 1).</td>
</tr>
<tr>
<td>HW12</td>
<td>2</td>
<td>Construct a GUI layout.</td>
</tr>
</tbody>
</table>
Table 5.2 (continued)– Description of homework learning objectives after week 5.

<table>
<thead>
<tr>
<th>Homework</th>
<th>LO</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>HW12</td>
<td>3</td>
<td>Write code to enable each component of the GUI layout to function appropriately.</td>
</tr>
<tr>
<td>HW12</td>
<td>4</td>
<td>Create a GUI flowchart from a problem description (Problem 2).</td>
</tr>
<tr>
<td>HW12</td>
<td>5</td>
<td>Identify and implement error checking in a GUI for enhanced usability.</td>
</tr>
<tr>
<td>HW12</td>
<td>6</td>
<td>Coding Standards.</td>
</tr>
<tr>
<td>HW12</td>
<td>7</td>
<td>Professional Habits.</td>
</tr>
<tr>
<td>HW13</td>
<td>1</td>
<td>Manually produce the equation of a regression line using the two-point method.</td>
</tr>
<tr>
<td>HW13</td>
<td>2</td>
<td>Compute slope intercept SSE SST and r-squared using least squares method (Problem 2).</td>
</tr>
<tr>
<td>HW13</td>
<td>3</td>
<td>Use MATLAB to plot a linear trend line and calculate the equation of the line (Problem 2).</td>
</tr>
<tr>
<td>HW13</td>
<td>4</td>
<td>Use Excel to plot a linear trend line over the raw data and display the equation and r-squared value.</td>
</tr>
<tr>
<td>HW13</td>
<td>5</td>
<td>Compute slope intercept SSE SST and r-squared using least squares method (Problem 3).</td>
</tr>
<tr>
<td>HW13</td>
<td>6</td>
<td>Use MATLAB to plot a linear trend line and calculate the equation of the line (Problem 3).</td>
</tr>
<tr>
<td>HW13</td>
<td>7</td>
<td>Coding Standards.</td>
</tr>
<tr>
<td>HW13</td>
<td>8</td>
<td>Professional Habits.</td>
</tr>
<tr>
<td>HW14</td>
<td>1</td>
<td>Plot data manually on linear and log scaled graphs.</td>
</tr>
<tr>
<td>HW14</td>
<td>2</td>
<td>Linearize a power exponential and/or logarithmic equation.</td>
</tr>
<tr>
<td>HW14</td>
<td>3</td>
<td>Create graphs in MATLAB with linear and/or log axis scales.</td>
</tr>
<tr>
<td>HW14</td>
<td>4</td>
<td>Use MATLAB to linearize a data set and plot linearized data.</td>
</tr>
<tr>
<td>HW14</td>
<td>5</td>
<td>Use linear or log axis graphs as a tool for identifying a function.</td>
</tr>
<tr>
<td>HW14</td>
<td>6</td>
<td>Create graphs in Excel with linear and/or log axis scales.</td>
</tr>
<tr>
<td>HW14</td>
<td>7</td>
<td>Coding Standards.</td>
</tr>
<tr>
<td>HW14</td>
<td>8</td>
<td>Professional Habits.</td>
</tr>
</tbody>
</table>

had negative correlations to some of the other learning objectives. There were two increases in the correlations among the learning objectives. First, correlations among learning objectives increased during week 4 (from HW04-5). After week 4 learning objectives had higher correlations to other learning objectives and course success than the first four weeks. Second, the correlations among learning objectives increased from week 9. Learning objectives from week 9 to 14 had higher correlations than previous weeks.
Figure 5.3 – Correlations among homework learning objectives (image best viewed in color).
5.3 Top predictive learning objectives

Feature selection methods highlighted learning objectives that are good predictors of student success in the course. In this section, I review the learning objectives that were identified by each of the three feature selection methods. The top five learning objectives are shown in Table 5.3 for the three feature selection methods. Correlation and Gini gain resulted in the same top four learning objectives. Explained variance had only one common variable with Gini gain and two common variables with correlation methods.

<table>
<thead>
<tr>
<th>Feature Selection</th>
<th>Learning Objective</th>
<th># of times appeared in top 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Correlations</td>
<td>HW05-6 - Professional Habits.</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>HW05-1 - Construct an appropriate function definition.</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>HW05-3 - Create a user-defined function using rules for writing functions.</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>HW05-4 - Create test cases to evaluate a user-defined function.</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>HW05-2 - Demonstrate how to create a MATLAB plot for technical presentation.</td>
<td>2</td>
</tr>
<tr>
<td>Gini gain</td>
<td>HW05-1 - Construct an appropriate function definition.</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>HW05-6 - Professional Habits.</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>HW05-3 - Create a user-defined function using rules for writing functions.</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>HW05-2 - Demonstrate how to create a MATLAB plot for technical presentation.</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>HW05-4 - Create test cases to evaluate a user-defined function.</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>HW05-3 - Create a user-defined function using rules for writing functions.</td>
<td>3</td>
</tr>
<tr>
<td>Explained Variance</td>
<td>HW03-7 - Professional habits.</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>HW02-3 - Interpret and evaluate logical statements.</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>HW05-4 - Create test cases to evaluate a user-defined function.</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>HW01-6 - Coding Standards.</td>
<td>1</td>
</tr>
</tbody>
</table>

All top five learning objectives that were selected by the correlation and Gini gain methods were from homework 5. Two of the learning objectives from
homework 5, HW05-3 (Create a user-defined function using rules for writing functions) and HW05-4 (Create test cases to evaluate a user-defined function), were selected by all three methods. In addition, “professional habits” showed up in all three methods (although from two different homeworks, 3 and 5).
CHAPTER 6 - DISCUSSION

The purpose of this study was to conduct a proof of concept and showcase how to create a course specific prediction model using learning objectives. In this chapter, I discuss the answers to the research questions including development of the prediction models, selection of student performance data, and achievement of learning objectives in the course. I also provide pedagogical implications, guidelines to create course specific prediction models, limitations of the study, and suggestions for future research.

6.1 R.Q. 1 - Best prediction modeling method

In this section, I discuss the findings related to R.Q. 1, finding the best predictive modeling method:

R.Q. 1 - Of six different predictive modeling methods, as well as a seventh hybrid or Ensemble method (consisting of three of the most successful individual methods, see section 3.4 for more details), which is the most successful at identifying at-risk students, based on specified in-semester student performance data? Why is this method the most successful? Why are the other methods less successful?

In this study, I created course specific prediction models to identify at-risk students at week 5 of the semester in a first-year engineering course with more than
1,600 students. After comparing seven prediction modeling methods, the more accurate ones for identifying at-risk students were the Naive Bayes Classifier (NBC) and an Ensemble model consisting of three models (NBC, Support Vector Machine, and K-Nearest Neighbor). Usually ensemble models have lower error in predicting future data than any of the individual models that make them up (Dietterich, 2000); however, in this study, NBC performed as well as the Ensemble model.

From a theoretical perspective, the classification error consists of bias and variance, and reducing one will result in increasing the other (Williams & Simoff, 2006). Bias is the error due to inaccurate assumptions in the learning algorithm (Hastie, Tibshirani, & Friedman, 2009). Simple models such as those created by the NBC method have higher bias. Variance is sensitivity of the model to small changes, typically due to noise, in the training data. More complex models such as those created by neural networks or decision tree have higher variance (Geman, Bienenstock, & Doursat, 1992).

In this study, the reasons the NBC model performed better than the other models were because of (1) NBC’s simple structure and, (2) the low number of at-risk students in the course. In this study, NBC calculated conditional probability of a student being at-risk based on the training dataset, and its structure was simpler than other methods. In addition, while the dataset contained a large number of students, a low percentage of students were at risk of failing. In the case of a low training sample size, a simple model with high bias and low variance such as NBC has an advantage over complex models with low bias and high variance such as KNN because the latter will overfit (Sammut & Webb, 2011). Overfitting may occur when
the model is complex and has a high variance. In this case, the model does not learn the underlying relations, instead it memorizes the data including the noise in the training data (Everitt & Skrondal, 2002). Thus an overfit model has high accuracy for training data and low accuracy for future test data.

Considering Logistic Regression modeling as the baseline method, it is possible to divide the modeling methods into two categories: the ones that performed better than Logistic Regression in identifying at-risk students, the Support Vector Machine (SVM) and the Naive Bayes Classifier (NBC), and the ones that performed worse than Logistic regression, the K-Nearest Neighbor (KNN), the Decision Tree (DT), and the Multi-Layer Perceptron (MLP) neural network. The methods in each category have two properties in common: bias and variance. As explained earlier, the prediction error is consisted of bias and variance, and decreasing one usually increases the other one (Hastie et al., 2009).

Typically, KNN, DT, and MLP modeling methods result in models with low bias and high variance (Moody, 1994; Olson & Delen, 2008). All these three models (KNN, DT, and MLP) can easily overfit to the training dataset when the training size is small (Domingos, 2012). These modeling methods require a large number of training cases to properly be trained. If the number of cases is not enough (e.g., low number of at-risk students in this study) these models cannot learn the underlying relations in the data and will just memorize the training dataset. Thus they will have a low accuracy on predicting other datasets. In contrast, NBC and SVM modeling methods typically result in models with high bias and low variance (Colas, 2009; Valentini & Dietterich, 2004). Thus, they perform with better accuracy than methods
with more complex structures when the sample size does not contain a large number of cases (e.g., number of at-risk students in this study). In summary, because the number of at-risk students in the datasets was low, the simple models with high bias and low variance (i.e., NBC and SVM) performed better than the models with low bias and high variance (i.e., KNN, DT, and MLP). For courses with high student enrollments and a high percentage of at-risk students, the models with more complex structures may perform better than the simple models.

Although it was expected that the Ensemble model could have a higher accuracy than the NBC model, the two models performed the same in predicting future data. The Ensemble model consisted of three methods: (1) NBC, the best method in predicting at-risk students, (2), KNN, the best method for predicting students who passed, and (3) SVM, the second best method in predicting at-risk students. Thus, SVM was designed to act as the swing vote and reduce the prediction errors. This design improved the accuracy of the predictions when all of the variables were used in the models. However, it did not improve the accuracy of predictions when only two or three variables were used in the models. In this case, SVM performed the same as NBC; thus, the result of the Ensemble and NBC models were the same regardless of the KNN predictions.

The main reason the SVM performance became similar to NBC in predicting at-risk students was selecting a only two variables by the feature selection methods instead of using them all in the models. In addition to size of the dataset, the SVM modeling error is related to the complexity of its structure, which is related to the number of predicting variables in the model. Decreasing the number of predicting
variables resulted in a simpler SVM model structure, which increased its bias and reduced its variance error. Because the number of at-risk students was low, SVM performed better with a lower number of predicting variables. This can also explain why the Ensemble model performance decreased more compared to NBC with the increase in predicting variables in the models. When the number of predicting variables in the model increased, the SVM model accuracy decreased, and depending on that, the Ensemble model accuracy decreased as well.

NBC is designed based on the naive assumption that the predictor variables are conditionally independent (Duda, Hart, & Stork, 2012). This assumption is used to simplify the calculation of conditional probability in the model. However, NBC typically performs well even if the predicting variables are correlated and this assumption is violated (Bishop, 2006). NBC performance may not decrease when there are correlations among predicting variables if the correlations, no matter how strong, are evenly distributed in categories (Zhang, 2005). The results of this study are another example of violation of the naive assumption, because a student’s grades are correlated to each other and not independent but NBC performed with high accuracy.

Even the more accurate models failed to identify some of the at-risk students. The models could not identify 33 (out of 147) students who were at-risk. Out of these 33 students, 22 students received a D grade, and only 11 students actually failed the course and received an F grade. While missing at-risk students has negative educational consequences, these numbers compared to the total number of students in the course (1413 students) are very low.
There are multiple reasons why the models could not identify some of the students correctly. First, as statistician George Box said: “All models are wrong but some are useful” (Box, 1979, p. 209). From a theoretical perspective, in case of identifying at-risk students, it is not possible to truly model students’ success because a lot of factors influence students’ behavior and success in a course. Any prediction model is an estimation of students’ performance and has its limitations. These limitations will cause the model to misidentify some of the students. In addition, students’ behaviors are not consistent throughout the semester. For example in this study, the models were trained based on the first five weeks of the semester. It is possible that a student has good performance at the beginning of the semester but begins to perform poorly in the middle or at the end of the semester or vice versa. In either situation, the models are not able to identify this student correctly.

Second, one of the challenging aspects of creating a prediction model for this first-year engineering course was the fact that fewer than 10% of the students were at-risk. All prediction modeling methods try to increase the overall accuracy of the predictions. Thus, it is more likely that the created model will perform better for categories with higher number of cases as compared to the categories with lower number of cases. Because there were a low number of failing students, most of the models had greater accuracy for students who passed than students who failed.

The third reason that the models were not able to identify all students correctly is the limitations of the models by pedagogical choices in the course. The final model used the midterm exam grade and one learning objective score. Several
factors including the extent to which exam and homework questions assess the
desired outcomes in the course, how good these assessments have been designed,
and the reliability of grading, influences these variables predictability of students' success. This ties back to the main goals of backward design (Wiggins & McTighe, 1998) and standards-based grading (Atwood & Siniawski, 2014). A well-defined learning objective with a well-designed assessment that is being graded based on a well-defined rubric not only increases fairness and transparency in the course (Sadler, 2005) but also increases the chance to provide a reliable grade with high predictive power for identifying at-risk students.

Fourth and last, in this course only homeworks were graded based on learning objectives, and midterm exam and quizzes were not. The midterm exam grade weight in determining the final grade was much higher than the individual learning objective scores. Thus, the midterm exam grade was always included in the prediction models. It is possible that not all of the exam questions were good predictors of students’ success in the course, but because they were bundled in one single grade, unlike homeworks that assessed 6-7 learning objectives, it was not possible to select the parts of the midterm exam that were better predictors of students’ success. Therefore, using course performance data including all assessments that are graded based on learning objectives may result in selection of a few midterm learning objectives that can increase the accuracy of the predictions.

This study is the next step in the development of early warning systems such as Course Signals (Pistilli & Arnold, 2010) that use generic prediction models. The development of accurate course-specific prediction models in this study confirms
scholars’ claims (e.g., Macfadyen & Dawson, 2010) that course-specific prediction models lead to higher accuracy of the predictions.

This study also expands on previous research by examining the class size range for which the prediction models can be used. In this study, the accuracy of the predictions reached their best performance after the training dataset contained at least 120 students. The performance of the models stayed almost the same from 120 to 780 students, which was the full training dataset. Thus it is not clear what the maximum boundary for training the models should be to avoid overtraining the models. Overtraining may occur with a very large training dataset. In case of overtraining, the model learns the training data too closely, which may prevent it from finding the underlying relationships (Nisbet, Elder IV, & Miner, 2009), and the model performs well for the training dataset but very poorly for other datasets (Zaki & Meira Jr, 2014).

The main reason the models’ performance did not drop with the increase in the size of the dataset is that 780 cases are typically not considered too many for any of the modeling methods especially when fewer than 10% of students (i.e., fewer than 78 students) belong to one of the categories. A larger dataset is needed to determine the maximum boundary of the modeling methods in identifying at-risk students.

In the training dataset, fewer than 10% of students failed the course, and a subset of 120 students had less than 12 students who were at-risk of failing the course. It is likely that the low performance of the models with datasets smaller than 120 was due to the limited number of at-risk students in the datasets. If a course has
a greater number of at-risk students, it may be possible to create prediction models for the course with less than 120 students.

6.2 R.Q. 2 – Selection of student performance data

In this section, I discuss the findings related to R.Q. 2 and its pedagogical implications:

R.Q. 2 - To what extent can the models created by predictive methods for identifying at-risk students in a course be improved through the selection of in-semester student performance data (e.g., quiz, homework learning objectives, midterm exam)? What does the selection reveal?

Using three feature selection methods (correlation, Gini gain, and explained variance), I identified the course assessments that had the greatest predictive power. In addition to the midterm-exam grade, homework 5 learning objectives were better predictors of students’ success than other learning objectives. The accuracy of predictions when using only two predictors (midterm exam grade and a single homework 5 learning objective) was 90.1%.

The main idea behind feature selection is that the data contains many features (i.e. variables), which may be redundant or irrelevant to the outcome variable, thus adding them in the prediction models will not improve the accuracy of the predictions (Guyon & Elisseeff, 2003). Using feature selection in predictive modeling or selecting a subset of predicting variables instead of using all of them has three main advantages. First, it reduces the models training time because models with fewer variables have fewer parameters to estimate. Second, models with a lower number of variables are simpler and easier to understand and
interpret (James et al., 2013). This is especially important from a pedagogical perspective because it helps educators understand what leads to students’ success in a course. Third, feature selection increases the models generalizability and their accuracy to predict future data by avoiding overfitting (Bermingham et al., 2015).

As it was expected, in this study, using feature selection methods increased the accuracy of predictions for future data. The models yielded similar results for identifying at-risk students in the current semester and future semester by employing feature selection methods. This was because using feature selection avoided overfitting to the training dataset and made it possible for the models to have similar accuracy for training and future data.

Although the primary purpose of predictive modeling is to predict an outcome variable, in educational settings it is always helpful to understand the models. In this study, understanding the prediction models has pedagogical benefits. By understanding the models, the instructor can map course assessments and learning objectives to students’ success in the course and identify the assessments and learning objectives that are not related to students’ success. For example, in this study and similar to Olama et al. (2014), quizzes are not as useful as homeworks in identifying at-risk students.

Olama et al. (2014) predicted students’ success in a course using homework, quiz, and participation in online discussions by employing neural network and logistic regression models. Similar to the current study, they found that homework grades are a better predictor of students’ success than quiz grades. They did not use midterm-exam grades in their model. From a pedagogical perspective, although
quizzes do not contribute to students' success in the course, they may have other value. Thus, use in a course should be carefully considered to make sure they have educational value for the students.

This study expands the findings of Huang and Fang (2013) and Olama et al.'s (2014) studies in terms of the types of prediction models evaluated and number of variables utilized in the models. Huang and Fang (2013) used only a mid-term exam grade in their models. The midterm exam grade had similar predictive power to pre-course performance indicators such as GPA and pre-requisite course grades. This study expands the results of Huang and Fang (2013) by comparing the predictability of midterm-exam grade and performance on weekly assessments including homeworks and quizzes. Midterm-exam grade, because of its high weight in the computation of students' final grade, was the best predictor of students' success. Homework learning objective scores were very similar to the midterm-exam grade as good predictors of students’ success. Thus, it is possible to use homework learning objective scores in early weeks of the semester instead of past performance or a midterm exam grade. This has two advantages. First, instructors do not need to wait until the middle of the semester, when a midterm exam is taken, to identify at-risk students. Second, using past performance in the prediction models may discourage students from working hard and doing their best in a course because they cannot change their past performance and may feel they are destined to fail no matter how hard they try.

Compared to learning objectives, quizzes did not have high predictive power and were not correlated to course success or even to each other. Quizzes were not
being graded based on Standards-Based Grading (SBG), and each instructor might have done something different for a quiz. This inconsistency in design and grading of the quizzes might be the reason for the low predictive power of quiz grades.

### 6.3 R.Q. 3 – Achievement of learning objectives

In this section, I discuss the findings related to R.Q. 3 and its pedagogical implications:

R.Q. 3 - What are the relationships, if any, between students’ success and achievement of different learning objectives in a course? What are the implications for the resulting prediction models and what are the pedagogical implications?

Use of learning objectives in the prediction models can help identify potential threshold learning objectives (Meyer & Land, 2013) that are critical to students’ success in the course. This adds to the benefits of utilizing SBG and increases transparency and fairness of grading (Sadler, 2005).

This study highlighted the usefulness of learning objectives in predicting students’ success in the course. In identifying important learning objectives in the first five weeks, week 5 leaning objectives were more important than the first four weeks. Most of the top five learning objectives selected by the feature selection methods were from week 5. One explanation may be that in week 5 students learn how to define user-defined functions in MATLAB and use that knowledge for the rest of the semester in their assignments. Thus, if a student does not learn how to do that, she will not be able to successfully complete later assignments in the course because they all require using user-defined functions. In addition to week 5, the
content of week 2, relational operators and logical statements, is also being used for the rest of the semester. However, week 2 learning objectives were not as important as week 5 in identifying at-risk students. One possible explanation is that the content of week 2 was easier to understand, and all students including those at-risk did well in this week.

The most important learning objective was “professional habits”. To achieve this learning objective, students must properly follow the assignment instructions; for example, name their files correctly. While this may seem trivial to students, for user-defined functions proper naming is necessary for the functions to work. In addition, students who paid attention to the instructions are more likely to spend more time on the course materials, learn the concepts, and do the assignments correctly.

Most of the top five learning objectives selected by feature selection methods were week 5 learning objectives. In addition, before this week, there is not a high correlation between success in the course and achievement of the learning objectives. After week 5, the correlations among learning objectives as well as learning objectives and course success increases. Thus, week 5 seems to be a critical week for students’ success. In addition to the critical content of this week, the first written exam was also after this week. Correlations of learning objectives and course success increased from week 5. Thus, it seems students who became successful in the course started to take the course seriously, latest at week 5 and spent more time and effort on the assignments from this week forward.
One possible explanation is that students start to study for the course close to the first written exam, which in this case was at week 5. One possible hypothesis is that having a written exam early in the semester may motivate students to start studying for the course early in the semester. Although there were weekly homeworks and quizzes in the course, students, if they ignore anything, are more likely to ignore them than a written exam. Thus, having an early written exam may increase students’ success in the course. I could not find any research study supporting this hypothesis and it needs to be investigated. If this hypothesis is confirmed, one way to increase predictive power of homework learning objectives in the course is by administrating a written exam early in the semester.

All students, even the ones who failed the course, did relatively well on homeworks 1, 2, and most of the homework 3 learning objectives. Homeworks 3 and 4 learning objectives were not related to the other parts of the course. In addition, successful students did not perform as well as the previous homeworks on homework 4. Thus, week 5 was the first week of the semester that the successful and at-risk students’ achievement on learning objectives started to differ.

One possible explanation is that even successful students ignored the first few weeks of the semester. Another possible explanation is that the course content for the first few weeks of the semester was not as important as the rest of the semester for success in the course, or they were easy enough that most of the students, regardless of passing or failing the course, could learn them. Thus, these weeks’ learning objectives did not have high correlations with course success and were not selected by the feature selection methods as important learning objectives.
Some quizzes and learning objectives were not related to course success. In addition, a few of the learning objectives were not correlated to other learning objectives even within the same homework. This indicates a problem either in students' understanding of the concept, homework or quiz design, or grading of these assessments. These results provide a chance for instructional designers and educators to identify the assessments that are not correlated with success in the course. This can be used as a diagnostic tool for engineering educators and instructional designer to find the flaws in the design of the course or its assessments.

Two of learning objectives at week 4, HW04-1 and HW04-4 (Interpret cumulative distribution plots for problem 1 and 2) had negative correlations with other learning objectives. Low correlations with success in the course and other learning objectives may indicate a pedagogical problem. This problem may be due to the design of the homework question or the grading rubric. For example in this case, the rubric had the following condition for achievement of the learning objective: “answers to parts B1 – B3 are correct and adequately justified.” One explanation for low reliability of the grades can be the vagueness of the rubric for the graders. In this case, it is possible that graders had different interpretations of “adequately justified”. This is an example of how use of learning objectives in prediction models can lead to investigations that inform the design of the assessments and highlight the ones that are poorly designed or graded.

This study expands the understating of SBG by identifying another advantage that was not explored before, which is the fact that SBG provides higher accuracy in
identifying at-risk students. This study used learning objectives scores in prediction models for the first time. First, learning objectives provide more data points than a single grade. Second, because learning objectives are graded based on well-defined rubrics (Heywood, 2014), in general SBG is more likely to provide reliable and consistent grading compared to grading without well-defined rubrics or criteria. Grading based on rubrics is especially beneficial in a course with multiple graders that have different levels of experience and education (i.e., instructors, graduate teaching assistants, undergraduate teaching assistants). With the benefits of utilizing learning objective scores in prediction models, it is highly recommended that engineering educators move toward SBG for all assessments in the course.

Creating prediction models as recommended in this study provides diagnostic tools for course instructors to learn about their course, find the weaknesses and strengths of the course, discover the potential threshold learning objectives, figure out whether or not the learning objectives, assessments, and course success are correlated to each other, and highlight the most and least important learning objectives for success in the course. Instructors can use this information to improve their course.

6.4 Recommendations for creating course specific models

This study took the first step in creating a course specific prediction model based on learning objective scores to identify at-risks students in a course. Based on the development of these models, in this section I provide recommendations for developing prediction models for researchers and instructors.
The minimum recommended class size to create a prediction model for a course with at least 10% at-risk students is 120 students. For a course with 120 students or more, it is possible to create a prediction model and identify at-risk students in the early weeks of a semester. For smaller size classes, in which it is not possible to train and utilize prediction models, an instructor’s experience and her familiarity with the students may be more useful than employing an inaccurate prediction model. The models are more useful for classes with a large number of students, in which the instructors are less likely to know which students are struggling during the semester. However, small class sizes, probably with fewer than 40 students, may not need prediction models. In a small class, the instructor has a chance to know the students and identify and help at-risk student without using a model.

Creating an ensemble model consisting of the best two prediction models for at-risk students, and the best model for identifying successful students will likely reduce the prediction errors and increase the accuracy of the predictions. In this study, the Naive Bays Classifier (NBC) was a suitable model for identifying at-risk students based on performance data in a course with a low percentage of at-risk students. In classes with a higher number of at-risk students and high enrollments, more complex prediction modeling methods such as KNN, DT, or MLP may provide higher accuracy than NBC and SVM. However, these models’ performances need to be investigated. The performance of prediction modeling methods with regard to class size and percentage of at-risk students is summarized in Table 6.1.
Table 6.1 – Guidelines for using prediction modeling methods.

<table>
<thead>
<tr>
<th># of students in the course</th>
<th>% of at-risk students</th>
<th>Know the students</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low</td>
<td>Low</td>
<td>NBC, SVM</td>
</tr>
<tr>
<td>High</td>
<td>High</td>
<td>KNN, DT, MLP*</td>
</tr>
</tbody>
</table>

* need to be investigated

Because it is not possible to create a model without any errors, it is important to communicate the results of identifying at-risk students with consideration to this fact. If a student performs well at the beginning of the semester, and then for any reason her performance drops, the models will not be able to identify her as an at-risk student. Thus, it is important to communicate the models' limitations in identifying at-risk students to students.

To create a prediction model, the course structure should not change drastically from one semester to another. One of the changes that can significantly influence the models prediction accuracy is changes to the learning objectives for the course. Since the models are trained based on the learning objectives and use them as predicting variables, change in the learning objectives results in invalidating the models. In such cases, the models need to be re-created and re-trained based on the new training data so that they can be used for identifying at-risk students. In this case, the feature selection methods may select different assessments and learning objectives that are good predictors of students' success in the course. This will change the structure of the models. The new models then need to be trained based on the new training data and tested based on the new test data. Thus, to train and test the models the data from at least two semesters is required.
From a pedagogical perspective, the new models can help the instructors learn if changes to the course are aligned with students’ success in the course or not. For example, it is possible to add a new form of assessment instead of a quiz to a course. Selection of this new assessment's grades by feature selection methods is a good indicator of their alignment with students’ success in the course. Change in grading of midterm exam to be based on learning objectives may result in selection of a subset of midterm exam learning objectives, which may lead to higher accuracy of the predictions. In this case, instructors can learn which questions in the midterm exam are related to success in the course.

Some changes in the course may not change the predicting variables but their predictive power. Examples of such changes are placing more emphasis on certain specific learning objectives in the course, changing the order of topics in the course, and increasing learning objective scores' reliability by redesigning the assessment questions and training the graders.

A large number of course-specific prediction models use only one semester's worth of data to train and test the models. The results of such models are optimistic and these models may not be useful for identifying at-risk students in future semesters. It is important to train and verify the models with more than one semester's data. It is necessary to train the model with one semester, and test its accuracy with another semester data to investigate how the model performs at identifying at-risk students in future semesters. If possible, it is useful to train the models with more than one semester. In this way, the models may adapt to changes
from one semester to another, and their accuracy for predicting at-risk students in future semesters may increase. However, this needs to be investigated.

Most of the course specific models only report their overall accuracy without optimizing false negative or type II error. Because the number of at-risk students is usually lower than the number of students who are successful in the course, most prediction models have higher accuracy for successful students than at-risk students, who need help the most. Thus, using an evaluation criterion, such as $F_{1.5}$, that evaluates the models in favor of at-risk students and helps reducing type II error is strongly recommended.

6.5 Limitations

One of the limitations of this study comes from difficulties in predicting human behavior. While the modeling methods identify patterns in data, it is not possible to make a 100% accurate model of students’ performance due to unpredictability of their behaviors and the various factors that influence students’ success in the course. Thus, the developed models provide only an estimation of students’ performance.

Other limitations of this study were due to some of the pedagogical decisions made by the course designer or instructor. For example, only homeworks were graded based on learning objectives. This decision limited the comparability between different forms of assessment in the course including quizzes, homeworks, and midterm exam. In addition, the fact that midterm exam was not graded based on learning objectives prevented identifying which sections of the midterm exam were more related to students’ success in the course.
Another limitation of this study was that some of the course assessments such as model-eliciting activities and design assignments were administrated after week 5. Thus, it was not possible to use them in the models for early prediction of at-risk students. These assessments may have good predictive power for identifying at-risk students.

The models depend on the performance data that is being collected during the semester. The quality of the course design and its assessments, and reliability of the grading are a few examples of pedagogical factors that influence the quality of the data and predictive power of the variables. If the performance data that is collected during the semester is not valid, it may not be possible to predict students’ performance at the end of the semester.

Another limitation is the necessary class size to train the models. Although a small class (e.g., smaller than 40 students) may not need a prediction model, the minimum class size for a course with about 10% at-risk students is 120. This study does not have a recommendation for identifying students at-risk in courses that are too big for the instructor to know the students and too small to build a prediction model (e.g., between 40-120).

The necessary underlying condition to develop a course-specific prediction model based on one semester and use it to identify at-risk students in another semester is that the course syllabus and structure should not change drastically from one semester to another. Thus, recommendations in this study only apply to courses that are fairly stable and do not change from one semester to another. For a
course with a lot of changes, generic prediction models may be used, however, they may have low accuracy.

If it were possible to run the models over an extended period or time (with multiple years worth of data), the results would become more stable and reliable. However, in practice, courses typically change from one semester to another, and the changes over a few years span would limit the usability of very old data for predicting performance in more recent semesters.

6.6 Future research

The next logical research step after development of the models is to investigate research questions that can lead to effective use of the prediction models in educational settings. In future research, it may be possible to find the optimal time to utilize the prediction models during the semester. Predictions made too early may not be very accurate, but late predictions may come when there is too little time to help students. In addition, how the prediction results should be communicated to the students can be investigated. For example, while email notifications are less time consuming for an instructional team, one-on-one meetings with the students may be more effective.

Other future research might investigate how often the models should be run and students informed about the results. Is it better to inform students weekly, monthly, or only once or twice during the semester? What course topics or learning objectives are most important for students’ success in a course? While this information may not be shared with the students, instructors can benefit from it and
focus on the possible threshold concepts and learning objectives to help students improve their performance.

In addition to predicting students’ success and identifying at-risk students at the end of the semester, it is also possible to predict students’ performance during the semester for example for written exams or projects during the semester. This can help students realize that they need to invest more time and effort to be successful in exams that usually have a high weight in the calculation of the course final grade. Thus, in addition to the overall threshold concepts and learning objectives in a course, it is possible to identify possible threshold concepts and learning objectives for each written exam in the course. The comparison of these threshold concepts to the learning objectives can be a valuable pedagogical tool for instructors to make sure all exams are aligned with each other and being successful in exams is aligned to being successful in the course.

Another way to help students become more successful is to learn what leads students to success in a course. Using descriptive learning analytics methods, it is possible to create profiles of successful and unsuccessful students in a course. Typically there is more than one path that leads to success. Similarly, not all students who fail a course have similar behaviors and problems. Identifying different student profiles for different courses can help both instructors and students recognize successful and unsuccessful behaviors.

Future research is also needed to investigate the minimum training size for courses with higher percentages of at-risk students. If a course has more at-risk students, it may be possible to create a prediction model with a training dataset of
fewer than 120 students. In contrast, if a course has a lower number of at-risk students, the training dataset may need to be larger than 120 students. It is also possible to predict students’ success in each section of the course to investigate if there are any differences in the accuracy of predictions between sections. If such a difference exists, investigating the reason for these differences can help create better prediction models and also improve students’ success in some of these sections.

One hypothesis that was offered as a possible explanation in this study and has not been investigated is that students do not take the course seriously before the first exam and start to study for the course from the time of first written exam. It is possible to investigate the influence of time of the midterm exam as well as number of written exams in a course on students’ success, and learn whether or not earlier or more frequent written exams in a course lead to more successful students in the course.

6.7 Conclusion

This study developed learning objective based prediction models to identify at-risk students in a first-year engineering course with a high number of students. After comparing seven different models using different numbers of variables, a simple model (Naive Bayes Classifier) with a low number of variables (midterm exam grade and one learning objective score) was considered to be the best model to identify at-risk students for future semesters. This study also identified new benefits of standards-based grading in engineering courses. Use of learning objectives in the models resulted in higher accuracy of the findings than other forms
of assessments. In addition, this study identified potential threshold concepts and learning objectives as well as the learning objectives that were not correlated with other learning objectives in the course.
LIST OF REFERENCES
LIST OF REFERENCES


Campbell, J. P. (2007). *Utilizing student data within the course management system to determine undergraduate student academic success: An exploratory study.* (PhD), Purdue University. Retrieved from [http://search.proquest.com/docview/304837810](http://search.proquest.com/docview/304837810)


VITA
VITA

Education & Certificates

Doctor of Philosophy (Ph.D.) in Engineering Education  
Purdue University, West Lafayette, Indiana  
[Jan. 2016]
• Dissertation: A Learning Objective-Based Model to Predict Students’ Success in a First-Year Engineering Course
• Advisory Committee: Dr. Diefes-Dux (chair), Dr. Madhavan (co-chair), Dr. Ohland, Dr. Main
• Graduate Certificate in Psychological Statistics (expected Dec. 2015)

Master of Art (M.A.) in Educational Technology and Learning Design  
Simon Fraser University, Vancouver, Canada  
[May 2012]
• Thesis: Design and Implementation of a Graphical Interface for Online Discussions
• Advisory Committee: Dr. Wise (chair), Dr. O’Neill, Dr. Kirkpatrick
• Certificates of Instructional Skills Workshop (ISW) and International Teaching Assistants (ITA)

Master of Science (M.S.) in Computer Engineering, Artificial Intelligence  
Amirkabir University of Technology, Tehran, Iran  
[Jan. 2005]
• Thesis: Recovery of Planar Parallax (3D image depth) from Multiple Frames
• Advisory Committee: Dr. Rahmati (chair), Dr. Safabakhsh, Dr. Kabir

Bachelor of Science (B.S.) in Computer Engineering, Software  
Sharif University of Technology, Tehran, Iran  
[June 2002]

Awards

• Duncan Fraser Best Student Paper Award at Research in Engineering Education Symposium, 2015
• Purdue Graduate Student Government Travel Grant, 2015, $1000
• Purdue Graduate Student Government Professional Grant, 2013 & 2014, $500 each
• Ross Fellowship, Purdue University, 2012, ~$52,000
• Graduate Fellowship, Simon Fraser University, 2011, $6,250

Grant Writing Experience

National Science Foundation (NSF EEC 1329304) grant, 2013
• Amount awarded: $300,000
• Title: Expert-Novice Framework to Support Student and Instructor Feedback on Design
• Helped design the research study and write the proposal

Discovery, Engagement and Learning (DEAL) Grant, Purdue University, 2012
• Amount awarded: $2,500
• Title: Investigating Graduate and Undergraduate TAs’ Perspectives Regarding Model-Eliciting Activity Implementation in the First-Year Engineering Program
• Wrote the proposal as a member in an interdisciplinary team of four graduate students
Educational Research and Teaching Experience

**Graduate Research Assistant**  
*Jan 2014 – Present*

School of Engineering Education, Purdue University  
Feedback on Design National Science Foundation Project (NSF EEC 1329304)
- Co-operated in writing the project proposal to submit to NSF
- Developed a coding scheme and analyzed student, instructor, and TA feedback data
- Conducted professional development feedback workshop for educators and systems engineers

**Ph.D. Student**  
*Jan 2012 – Present*

School of Engineering Education, Purdue University
- Used learning analytics to identify at-risk students in a first-year engineering course (dissertation)
  - Analysed more than 3300 students grades (~300,000 records)
  - Implemented neural network, decision tree, support vector machine (SVM), k-nearest neighbour, and naïve bayes classifier prediction models
- Implemented naïve bays classifier and k-means clustering in python for a computer science data mining course
- Investigated influence of time of class on students’ grades in an active learning course with HML analysis
- Evaluated grading reliability for model-eliciting activities in first-year engineering (FYE) program

**Graduate Researcher**  
*Apr 2012 – Apr 2014*

DEAL Project, Purdue Graduate Student Government (PGSG), Purdue University
- Prepared research proposal and IRB application
- Prepared and conducted semi-structured interviews to understand FYE TAs’ perspectives
- Created and administered a survey instrument
- Analyzed interview and survey data

**Graduate Teaching Assistant of Preparing Future Faculty/Professionals**  
*Jan 2013 – Dec 2013*

Graduate School, Purdue University
- Help the course instructor prepare reading materials
- Graded assignments and managed the course on Blackboard
- Designed rubrics for evaluating faculty candidates portfolios
- Administered a self-assessment survey to assess students’ employability skills

**Graduate Research Assistant**  
*Jan 2012 – Dec 2012*

INSPIRE, Purdue University
- Developed a prototype of SMART board application to support Model-Eliciting Activities

**Master Student**  
*Sep 2009 – Dec 2011*

Simon Fraser University, Vancouver, Canada
- Designed, implemented, and tested a new graphical interface to support online discussion forums interactions
- Co-developed “Live Learning Poll” Moodle block for formative evaluation during online and face-to-face classes (published on www.moodle.org)
- Designed multimedia learning modules (Map of Canada and Multiplication of signed numbers)

**Research Assistant**  
*Oct 2009 – Dec 2011*

Simon Fraser University, Vancouver, Canada
- Cooperated in a three year study on students’ participation in online discussion forums including literature review, data gathering, data analysis, and writing reports and papers
- Modified a PHP & MySQL online discussion forum (Phorum) in order to log click stream data and have a particular interface and functionality required for the study
• Analyzed more than 100,000 data logs using SPSS and MS Excel macro programming
• Developed a survey on students participation in online forums
• Conducted micro-analytic case studies to understand student behaviors in online forums

**Faculty Member**
Jondi-Shapoor University of Technology, Iran
*Sep 2005 – Aug 2009*
• Started undergraduate computer engineering program and designed its curriculum
• Instructed computer courses including Computer Programming (C++), Data Structures, Numerical Methods, and MATLAB workshop
• Conducted research on e-learning and published two papers in national conferences

**Professional/Leadership Experience**

**Director of Research**, Jondi-Shapoor University of Technology, Iran  
*Sep 2007 – Aug 2008*
• Appointed by the university president and worked in directly under the university vice-president
• Oversaw disbursement of internal research grant funds
• Updated university research policies to encourage faculty members to conduct more research

**Technical Manager**, Electronic Technologies Development Center, Iran  
*Jan 2005 – Aug 2007*
• Installed and administrated web based systems (e.g., Moodle and Joomla) for online courses and seminars
• Administrated Tehran Regional Electrical Company LAN and WAN including more than 20 servers and 600 workstations
• Organized ICT Marketing 2007 Seminar and Electronic Government 2006 Congress

**E-learning Technologist**, International University of Iran, Iran  
*Jan 2004 – Dec 2004*
• Installed, configured, customized, and administrated Moodle
• Assisted teachers in lesson design and learning content creation

**Technical Staff**, Sharif University of Technology, Advanced ICT Center, Iran  
*Sep 2002 – Dec 2003*
• Organized the First Iranian Linux User's Seminar
• Created a virtual class for Kish University with real-time video streaming
• Built educational multimedia CDs for Payame Noor, a distance learning University
• Administered Linux and FreeBSD servers

**Peer Reviewed Journal Publications**


**Peer Reviewed Conference Publications/Presentations**

13. **Marbouti, F., Diefes-Dux, H., & Strobel, J.** (2014). You may be able to teach early classes, but students may not be awake yet! *Proceedings of the 2014 American Society for Engineering Education Annual Conference*, Indianapolis, IN.


Workshops/Invited Talks

1. Rodgers K. & Marbouti, F. (November, 2015). *I wish I knew ... about research!* Professional Development workshop presented to engineering education undergraduate researchers, West Lafayette, IN.
3. Marbouti, F., & Rodgers, K. (August, 2015). *How to give constructive feedback?* Professional development workshop for teaching assistants of first-year engineering honors courses, Purdue University, West Lafayette, IN.
5. Wise, A. F., & Marbouti, F. (September, 2011). *Supporting “listening” in online discussions: A visual interface to promote engagement with chains of ideas.* Invited presentation given at the Socializing Intelligence through Academic Talk and Dialogue Conference hosted by the Learning Research and Development Center, University of Pittsburgh, Pittsburgh, PA.

Service/Volunteer Experience

**Conference paper reviewer**
- American Society for Engineering Education (ASSE) annual conference, 2015
- Frontiers in Education (FIE), 2013 & 2014
- First-Year Engineering Experience (FYEE), 2012

**Member of ENE Graduate Student Association Social Committee**  
Sep 2012 – May 2014
Purdue University
- Organized hiking trips for engineering education graduate students to encourage a supportive community

**Mentor,** International Mentorship Program, Simon Fraser University, Canada  
Aug 2010 – Aug 2011
- Helped new international students adapt living in Canada to ensure a more positive environment

**Web Programmer,** Drug Abuse Treatment N.G.O., Iran  
May 2009
- Developed a web based (PHP & MySQL) content management system used as the NGO’s central database

**Technician,** Sharif University of Technology, Computer Engineering Department  
Oct 2000 – Sep 2001
- Administered Linux and Windows Servers
- Supported Local Network and PC Workstations (Hardware and Software)

**Memberships**
- Member of Society for Learning Analytics Research (SoLAR)
- Member of American Society for Engineering Education (ASEE)
- Member of American Educational Research Association (AERA)