2015

A Predictive Modeling System: Early identification of students at-risk enrolled in online learning programs

Mary L. Fonti
Nova Southeastern University, fonti@nova.edu

A Predictive Modeling System: Early Identification of Students At-Risk Enrolled in Online Learning Programs

by

Mary Fonti

A dissertation submitted in fulfillment of the requirements
for the degree of Doctor of Philosophy

College of Engineering and Computing
Nova Southeastern University

2015

We hereby certify that this dissertation, submitted by Mary Fonti, conforms to acceptable standards and is fully adequate in scope and quality to fulfill the dissertation requirements for the degree of Doctor of Philosophy.

Steven R. Terrell, Ph.D.
Chairperson of Dissertation Committee

12-4-15
Date

Sumitra Mukherjee, Ph.D.
Dissertation Committee Member

12/4/15
Date

Thomas MacFarland, Ph.D.
Dissertation Committee Member

12-4-15
Date

Approved:

Amon B. Seagull, Ph.D.
Interim Dean, College of Engineering and Computing

12-445
Date

College of Engineering and Computing
Nova Southeastern University

2015

Table of Contents

List of Tables

List of Figures

A Dissertation Paper Submitted to Nova Southeastern University
in Fulfillment of the Requirements for the Degree of Doctor of Philosophy

Mary Fonti

December 4, 2015

Predictive statistical modeling shows promise in accurately predicting academic performance for students enrolled in online programs. This approach has proven effective in accurately identifying students who are at-risk enabling instructors to provide instructional intervention. While the potential benefits of statistical modeling is significant, implementations have proven to be complex, costly, and difficult to maintain. To address these issues, the purpose of this study is to develop a fully integrated, automated predictive modeling system (PMS) that is flexible, easy to use, and portable to identify students who are potentially at-risk for not succeeding in a course they are currently enrolled in. Dynamic and static variables from a student system (edX) will be analyzed to predict academic performance of an individual student or entire class. The PMS model framework will include development of an open-source Web application, application programming interface (API), and SQL reporting services (SSRS). The model is based on knowledge discovery database (KDD) approach utilizing inductive logic programming language (ILP) to analyze student data. This alternative approach for predicting academic performance has several unique advantages over current predictive modeling techniques in use and is a promising new direction in educational research.

Chapter 1

Introduction

The number of students taking online courses in the U.S. has increased by 570,000 between 2011 and 2012 totaling 6.7 million students (Blair, 2013). Although the average annual growth rate declined in the past few years to 9.3%, the proportion of all students taking at least one online course is at an all-time high of 32.0% (Allen & Seaman, 2013). As universities continue to experience tremendous growth in enrollment, retaining students and assuring academic success is a challenge for many institutions. While the average retention rate for traditional undergraduate students is 54% at private for-profit institutions and between 59% and 61% for private nonprofit and public institutions, it is significantly lower for students enrolled in online programs (Barber & Sharkey, 2012; NCES, 2013). As a result, the Department of Education (2014) has instituted new criteria addressing low retention rates. Institutions are required to demonstrate a commitment to academic success by redefining goals, collecting and analyzing information on student retention, making improvements based upon the analyses of student data, and implementing processes and methodologies for monitoring student progress (Higher Learning Commission, 2012).

Currently, the standard definition of retention established by the Federal Government and adopted by national accreditation commissions defines retention at the institutional level. Retention is defined and measured based upon students who are enrolled full-time in a degree program who remains enrolled from one fall semester to the next fall semester (HLC, 2012). The period of time tracked is typically six years for a four year college and three years for a two year college. Tracking and reporting to regional accreditation commissions is typically derived from data collected from first-time, traditional students. However, this does not include sub-populations of students who attend college part-time including students enrolled in distance education programs. These sub-populations are typically non-traditional students where retention rates are significantly lower. The Southern Association of Colleges and Schools Commission on Colleges (2011) and the Higher Learning Commission (2012) encourages institutions to choose measures that are suitable to student populations that are not included in criteria for accreditation. The SACSCOC and HLC also recommends conducting sustained, evidence-based and participatory inquiry which include documented assessment of student achievement conducted in each course by comparing student performance to the intended learning outcomes. Based upon these recommendations, a number of institutions have taken the initiative to create customized tracking to identify students who are at-risk at the course level by developing tools that serves as an early alert system.

In recent years, a number of research groups have begun to utilize machine learning techniques with a goal of improving retention rates by predicting academic performance with varying degrees of success (e.g., Agudo-Peregrina, Hernandez-Garcia, & Iglesisas-Pradas, 2012; Barber & Sharkey, 2012; Lauria, Baron, Devireddy, Sundararaju, & Jayaprakash, 2012). Machine learning techniques can handle analyses of large datasets making it feasible to develop analysis tools to better predict the correlation between factors impacting retention and associated outcomes (Luu, Rusu, Walter, Linard, Poidevin, Ripp, & Nguyen, 2012). Predictive models with associated statistical learning algorithms include Support Vector Machines (SVM), neural networks, decision tree or Naïve Bayes. Although these learning algorithms perform well for classification purposes, drawbacks include large memory requirements, lengthy computation times to deal with large datasets, and transformation of data into a standalone analytical software package for analysis (Luu, et al., 2013).

In response to these issues, a number of universities have developed integrated predictive systems with the goal of incorporating all functions and features within the educational system to improve the efficiency of access, maintenance and analysis. These integrated systems can be traced to Purdue University through the implementation of Course Signals. The system was pioneered by Campbell, Deblois, and Oblinger (2007) and implemented by Arnold and Pistilli (2012). Since the initiation of the first pilot project, retention rates have improved significantly. The premise behind the project was to develop an automated predictive modeling tool that integrated into Purdue's educational system. In a similar project conducted by Lauria, Baron, Devireddy, Sundararaju, & Jayaprakash (2012), the focus of the project was to expand on Purdue's Course Signals with the objective to develop an open-source model that is portable across a number of state-wide university systems. Results from both projects report between 82% to 90% accuracy in predicting academic performance. Limitations to the pilot projects include reliance on API extensions to transfer and transform data to facilitate the execution of algorithms for analysis potentially impacting performance and usability.

Predictive modeling is an emerging alternative to current predictive systems combining model and data management functionality to support user applications, analysis and system applications as a unified framework. Knowledge discovery from database (KDD) utilizing inductive logic programming (ILP) to automatically extract background knowledge to predict outcomes based upon inferred rules have been applied in a number of domains (Nguyen, Luu, Poch, & Thompson, 2013). KDD is defined as the process of identifying valid and understandable patterns in data (Džeroski, 2003). The KDD process involves selection and preparation of data, data mining and interpretations of the extracted results (Nguyen, et al., 2013). While the majority of machine learning algorithms accept as input a single table, this has led to the exploitation of logic reasoning approaches such as ILP by which a computer language can learn rules by example by extracting and analyzing data from multiple tables within existing database systems (Dzeroski, Cussens, & Manandhar, 2000; Fürnkranz, Gamberger, & Lavrač, N. 2012).

ILP relies on the theory of logic programming concerning semantics, inference rules, and execution. Based upon its rich representational language, prediction in the form of computational logic employs background knowledge in the induction process (de Raedt, 1998). Systems developed with ILP can learn a single concept (hypothesis) or multiple concepts (hypotheses) and accepts examples one by one (e.g. incremental learners) in the form of clausal formulas which can be revised in the learning process (Dzeroski, 2003). The main advantage of ILP over other machine learning algorithms is the learned patterns are expressed in symbolic form, which is easily interpreted allowing the integration of prior knowledge as part of the solution to the problem. This handcrafted rule approach can provide a complete and consistent view of all significant patterns in the data at the level of abstraction specified by the knowledge engineer (Lima, Oliveira, Pentagrossa & Freitas, 2013).

Problem Statement

As universities continue to experience tremendous growth in online courses, increasing enrollment has been overshadowed by low retention rates. While the average retention rate is 55% for traditional on-campus programs, it is significantly lower for online programs (Barber, et. al., 2012). As a result, regional accreditation agencies, who assure quality education to students, have instituted stricter standards requiring systematic tracking methods to monitor student progress.

With these recent demands and increased enrollment, instructors are faced with the challenge of closely monitoring student progress and providing support and resources. However, current learning management systems (LMS) do not provide instructors with effective tools that provide a comprehensive view of a student's academic performance early in the progress of a course. The traditional summative approach to evaluate and identify students who may be at-risk is provided at a stage in course progression where intervention strategies are ineffective (Macfayden & Dawson, 2010). Typically instructors have to wait until mid-term exams are completed to identify students who are at-risk (Huang & Fang, 2013).

Despite the growing number of studies focused on retention at the institutional level, development of viable tools to identify students who may be at-risk early at the course level is fragmented and lacking structure in the research field. Although statistical modeling techniques show promise in accurately predicting outcomes in a number of industries and fields, the number of studies investigating how higher education can benefit from applying these techniques is limited. Presently, machine learning techniques which have proven to accurately predict student outcomes are not well supported by relational database management systems (RDBMS) and software applications. Processes are affected by economic utility such as the cost associated with extraction of training data, transformation of data and model management (Guo & Paquet, 2013). Methods to rank or prioritize variables according to their predictive power utilizing various heuristic methods is a key ingredient missing in a number of studies using machine learning techniques (Lee & Shatkay, 2006). Additional drawbacks include handling multicollinearity, high error rates, outliers, and missing values (Freckleton, 2011; Fürnkranz, Gamberger & Lavrač, 2012).

As such, model-based prediction is emerging as an alternative method to standard predictive models. This paper contends that next generation database and software systems should natively support and manage predictive models, tightly integrating front-end Web processing, application programming interfaces, query processing of multi-relational databases and reporting. Exploiting predictive functionality within the institution's relational database management system is the natural progression that goes beyond current approaches.

Research Goal

The purpose of this experimental study is to develop an automated, Web-based predictive modeling system (PMS) that can be easily incorporated within a standard educational database management system. The PMS model will enable instructors to identify students who may be at-risk early in the semester. Although the system will be designed to run based upon default attributes, instructors will have the ability to manually select factors that are unique to type of course, program or student population. While prediction processes and procedures will be internal to the system, the Web-interface will be easy to use by instructors even if they do not possess a rudimentary understanding of prediction methods.

The benefits of the proposed PMS model includes easier integration into student-based systems and applications, the ability to process all functions utilizing an application interface and performing analysis by applying less complex computational formulas based upon knowledge discovery database (KDD) approaches using an inductive logic programming (ILP) approach that is easy to understand and to maintain. The predictive accuracy of the system should be comparable to those obtained using extant methods. The objective is to improve retention in online courses by providing instructors with a tool that performs real-time detection of students who may be at-risk for not successfully completing a course.

Research Questions

Research question 1: Among the selected combinations of academic and learning management system factors selected from the Web-based tool, which combination of factors accurately predicts student outcomes?

Research question 2: Is the Web-based predictive modeling system tool useful and easy to use when extracting, analyzing, and reporting student outcomes?

Research question 3: Is the predictive modeling tool a valid and reliable instrument for predicting student outcomes and monitoring student progress?

Research question 4: How easy is it for instructors to modify, maintain and manage the predictive modeling system?

Relevance and Significance

When students enrolled in online programs do not succeed, it comes at a high cost to the student, department and the institution they attend (Terrell, Snyder & Dringus, 2009). Costs are incurred in respect to time, resources and finances for students, faculty, institutions, and funding sources (Schneider & Yin, 2011). According to the American Institute of Research (2009), state and Federal Government lose approximately four billion dollars annually for the cost of students dropping out of community colleges. This number increases dramatically when you include students at the graduate level and students attending four year colleges.

For the non-traditional student, distance education provides increased access to new career opportunities. Often the student is unable to enroll in traditional on-campus programs due to employment and family obligations. As such, online courses provide a flexible and convenient opportunity to obtain a degree that would otherwise be unobtainable. For millions of students who are unemployed, dislocated, or displaced, online education provides viable options while seeking new employment (Betts & Lynch, 2009). The U.S. Census Bureau (2012), reports that individuals with just a high school diploma will earn on average $26,000 less per year than individuals with a bachelor degree. To date, only 30% of the U.S. population 25 years or older holds a bachelor or graduate degree. It is estimated that 63% of all jobs will require a degree by 2018 and there will be a shortage of 16 million college educated adults in the workforce by 2025 (Nunley, 2007). While college retention rates are improving in almost every post-industrialized country in the world, this is not the case for U.S. colleges and universities. As a result, college attainment is becoming increasingly important to the U.S. in order to compete in a competitive global workforce.

One of the key elements to improve retention rates for non-traditional student populations is accurately identifying students who may be at-risk for not succeeding in a course they are currently enrolled in. By doing so, students can be provided with the necessary resources and support to complete the course successfully. With a comprehensive view of a student's progress in real-time, instructors can have the opportunity to increase perceptions of support through feedback and social presence in the online environment (Park & Choi, 2009). Improved feelings of student-to-faculty connectedness by having an active and encouraging faculty presence is viewed as a contributing factor to improve persistence for students who would otherwise not succeed in an online course (Liu, Gomez, & Yen, 2009; Park & Choi, 2009; Terrell, Snyder & Dringus, 2009). Although non-academic issues such as work, family responsibilities, bereavement, and illness may contribute to a student not succeeding in a course, several studies reveal this can also be mitigated by the presence of strong support from faculty, staff and administrators (Aragon & Johnson, 2008; Bunn, 2004; Ivankova & Stick, 2007).

Barriers and Issues

While the proposed PMS system can expand upon current predictive modeling research, core challenges exist to effectively develop, implement and deploy said system. Understanding and addressing these issues prior to development of the model is vital to improving the chance of success.

Developing a Web-based tool that is compatible with a standard enterprise system and demonstrating this in a unified view is one of the key objectives of this study. With a wide range of disparities in technologies, data structure and applications along with fundamental differences in system architectures, these variances will need to be considered throughout the design of the system. The physical architecture and different software elements of the PMS as well as their characteristics must be defined. Definitions must be precise and use unambiguous language so that researchers and other stakeholders are left with no doubt as to the interpretation and rationale behind the selection of components and underlying concepts.

Establishing standards and addressing risk will provide a foundation for the study. Standards such as usability, reliability, performance, conformance, aesthetics, maintainability, and quality metrics must be met. Identifying risk is also vital. Risks may include but not be limited to constraints such as scope, schedule, quality, compatibility, and resources. Risk can also arise from unexpected problems or issues with projected estimates, assumptions or having limited information. It is important to understand and plan for issues that may occur and how it may impact the project and its objectives. Strategies to address and respond to risk minimizes the probability of project failure (Marchewka, 2006).

Assumptions

According to Leedy and Ormrod (2010), assumptions are so basic that without
an assumption, the research problem could not exist. In order to progress, it is important to
justify
why each assumption is true.

Assumption 1: An inductive logic programming (ILP) approach is a more efficient and effective method to predict if a student is at-risk for not successfully completing an online course.

Predictive modeling approaches in current educational research does not exploit the representational advantages of logic-based techniques to predict if a student is at-risk. Statistical relational learning (SRL), a sub-discipline of artificial intelligence (AI), is concerned with how a model in the domain handles both uncertainty and complex relational structures. Knowledge representation developed in SRL uses a subset of inductive logic programming (ILP) to deal with hypothesized predicate (propositional variable) definitions (Milch & Russell, 2006). Logic programming is differentiated from most other forms of machine learning (ML) techniques by its use of an expressive language and its ability to make use of logically encoded background knowledge. It is well suited for analysis of multi-relational datasets which is easily embedded, interpreted and maintained within a RDBMS.

Procedures, processes, and results from logic based predictive models is easy to interpret. However, predictive studies utilizing ML algorithms such as decision trees, neural networks and SVMs have poor interpretability and are often too complex to replicate. Establishment of training instances, classification, analyses and reporting of results is not easily understood. Predictive results are represented in standard graph form. Experts in the field of statistics are required to translate results into a more intelligible form.

Assumption 2: ILP approach handles missing fields more reliably than standard machine learning (ML) techniques.

Multiple imputation (MI) method in ML techniques is a standard approach used when dealing with missing values. This method has the potential for causing bias by using median, mean, or mode to populate fields (variables) that are missing. Although this is a preferred method now available in third party statistical software, it is a computationally intensive method that needs to be applied carefully to avoid misleading conclusions. Although deletion of records with missing values can increase variance and impact sample size, ILP enables one to logically exclude a record if multiple correlated factors are missing or a highly predictive factor is not available in a student record.

Assumption 3: ILP has equivalent predictive accuracy compared to standard ML statistical analysis techniques.

Unlike the majority of ML approaches currently in use, logic programming handles positive and negative training asymmetrically, focusing on inducing rules that match many positive examples and few (ideally zero) negative examples from multiple datasets (Kuusisto, Dutray, Nassif, Wu, Klein, Neuman, Shavlik & Burnside, 2013). A quantitative assessment using cross-validation of factors or a set of factors for significance can establish and prioritize the positive examples while eliminating the negative examples.

Limitations

There are a number of limitations which have the potential to impact the internal validity of the proposed study. First, developing a predictive model in a short time frame may impact the quality of the model. Additional time may also be required to replicate the study for a variety of online courses with different student factors and populations. Replication establishes the generalizability of findings while improving the confidence in regards to the reliability of the model. Secondly, there is a potential that variables that have not been considered for the PMS model may be important factors when predicting outcomes. No matter how extensive the research is expended to select variables, their still remains a degree of uncertainty as to which variables or combination of variables have the most predictive power. Finally, as historical data grows over time, it becomes more difficult to revise knowledge that accounts for new or changing theories and empirical evidence.

Delimitations

The student population investigated for this study will be delimited to students enrolled in an open, online course who are considered non-traditional based upon ages of 24 and over. While the goal of the study is to improve low retention rates of non-traditional students enrolled in online courses, the decision to select students who are enrolled in an open, certified, post-secondary course has a two-fold purpose: availability of data and a larger course sample size required to validate the PMS model. The non-traditional student population under investigation will be further bounded by the following available attributes for analyses: level of education, enrollment status, delayed enrollment in years (determined by date of birth and course registration date), gender, initial date of interaction with course compared to registration date, frequency of days active in the course, number of interactions with the video component, number of interactions within the courseware module, and number of chapters (assignments) completed at the end of the course. Variable selection is based upon established theory and seminal works examining characteristics of non-traditional students. Selection is also guided by the National Center for Educational Statistics (2013).

Definition of Terms

Algorithm is a list of well-defined instructions of computations that produce an output based upon selected input.

Cross-validation is a statistical technique used for estimating the performance of the PMS model.

Delayed enrollment is the measure in years a student graduated from high-school and enrolled in a post-secondary program.

Dynamic variable is a factor that is subject to change over time (e.g. marital status, dependents, post frequency).

Knowledge discovery in databases is the process of discovering useful knowledge and patterns from the population under investigation.

Logic reasoning refers to inductive reasoning that supports rules based upon established theory and extant literature to confirm (test) the hypothesized outcome.

Machine learning techniques refers to supervised learning models with associated algorithms used for classification. For this study, standard machine learning techniques refers to predictive retention studies that utilize: Support Vector Machines (SVM) which is concerned with mapping input (variables) into a higher dimensional space for classification purposes. Naïve Bayes, a highly scalable method, is based on linear time that requires less time to train and test and neural networks which estimates linear or non-linear functions minimizes cost criterion and employs a gradient descent.

MIT represents Massachusetts Institute of Technology.

Open-source describes how the code supporting the Web tool / PMS model is openly available for viewing or use within the research community for purposes of testing or improving functionality.

Non-traditional student is based upon definitions provided by NCES (2013). The non-traditional student is described as student 24 years or older that is enrolled part-time. The student is financially independent and typically has dependents and family obligations. Enrollment is delayed by a number of years between high school and enrollment in a college program / course.

Predictive statistical modeling for this study is concerned with the development of a model which forecasts a student final grade utilizing statistical techniques to validate the reliability of the models performance.

Retention for this study refers to non-traditional students who are enrolled in a certified online course at the undergraduate level after the course census date. If the student earns a satisfactory grade of 55% or higher the student receives a certificate from MIT and is considered successfully retained throughout the duration of the course.

Static variables is instantiated once and will remain constant throughout the course of the study (e.g. age, class level, GPA).

STEM refers to courses in science, technology, engineering and mathematics.

Variables consist of input (independent) items that predict the value of the output (dependent) or target item. In terms of this study, variables may be used interchangeably with the following: *attributes, characteristics, values or factors.* These terms change based upon context.

Acronyms

AI - artificial intelligence

API - application programming interface

CGI – common gateway interface

CPM – composite persistence model

CRN – course registration number

CV – cross-validation

DB – data base

EdX – Harvard / M.I.T. universities open, online learning management system

GED – General Education Diploma

GPA – grade point average

HLC – Higher Learning Commission

HTTP – hypertext transfer protocol

ILP – Inductive Logic Programming

INSTRID – instructor identification (renumbered – anonymous / unidentifiable)

LMS – learning management system

KDD – knowledge discovery in databases

MI – multiple imputation

MIT – Massachusetts Institute of Technology

ML – machine learning

MOOC – massive online open course

MIT/6.002 – MIT’s online Circuits and Electronics course

NSF – National Science Foundation

PDO – PHP data object extension interface for DB access to SQL server

PHP – hypertext preprocessor server side scripting language interface used for CGI

PLS – project life cycle

PMS – predictive modeling system

RDBMS – relational database management system

SACSCOC – Southern Association of Colleges and Schools Commission on Colleges

SDLC – system development life cycle

SID – student identification number (renumbered – anonymous / unidentifiable)

SME – subject matter expert

SQL – structured query language or server

SRL – statistical relational learning

SSRS – SQL server reporting services

STEM - Science, Technology, Engineering and Mathematics fields

W3C – World Wide Web Consortium

Summary

This experimental study expands on current research by proposing an alternative approach to predictive modeling. This paper contends that relational database management systems (RDBMS) should natively support all predictive functions and features by tightly integrating front-end Web processing, application programming interfaces, extraction, analysis, and reporting services utilizing ILP in conjunction with SQL query language. This method has several advantages. First, it can support predictive analytics to answer complex questions involving missing values, correlations and variable ranking. Secondly, data can be extracted and analyzed from a RDBMS improving workflow and reducing data transfer and transformation overhead. Most importantly, ILP enables the expression of conditions in computational logic based upon theoretical and background knowledge. From an operational point of view, all processes for this study will be designed to function within a standard educational system from variable selection, data preparation, analysis to final interpretation. This approach for predicting academic outcomes has several unique advantages over current predictive modeling methods.

Chapter 2

Review of the Literature

In the past decade, there has been a significant increase in enrollment in online programs at the post-secondary level. Despite continuous growth, one of the largest challenges for educational leaders is that student retention rates are significantly lower than traditional, campus-based programs (Allen & Seaman, 2013). Non-traditional students are the largest subset of students studying in online learning environments (OLE). Identification of students based upon factors unique to this population is crucial for improving retention rates (Shapiro, Dundar, Chen, Ziskin, Park, Torres, & Chiang, 2012). A handful of predictive modeling systems have been implemented at universities. They have shown promise for successfully identifying students who are at-risk.

This chapter provides an historical overview of student retention, theoretical perspectives of retention, characteristics unique to non-traditional students enrolled in online programs, predictive modeling techniques employed in current research to identify students at-risk, as well as barriers and issues that needs to be addressed to successfully implement predictive systems within standard educational systems. This chapter concludes with a summary of key findings that will contribute to the proposed research study.

1. Historical Overview of Retention

Early research investigating student retention in post-secondary education can be traced to a seminal study conducted by John McNeely in 1936. The author’s goal was to examine if specific demographic, institutional and social factors contributed to students not successfully completing a program of study. This longitudinal study followed 15,535 undergraduate students from 60 colleges and universities entering their freshman year through their senior year. McNeely identified several factors influencing retention. Results of the study revealed retention rates for freshman students was 33.8%. This was significantly lower than the overall average rate of 45.2%. Differences in gender, type of institution, college major and extra-curricular activities were found to influence whether a student persists and graduates from college. Although, McNeely's work was highly influential in laying the groundwork for future research, studies examining retention in higher education were limited between the time this study was published until the 1960s (Berger & Lyon, 2005). It was the general consensus that if a student did not successfully complete a program, the student was unqualified to achieve academic success at this level. Students failed, not the institution they attended (Tinto, 2006).

This view of retention began to shift in the late 1960s with a rapid growth in higher education. The G.I. Bill, Civil Rights Movement and the Higher Education Act (HEA) resulted in greater access to a diverse population of students. Students from lower and middle income households were provided financial support to enroll in college (McDonough & Fann, 2007). With increased enrollment, researchers began to examine the role external factors played in a student's decision to stay or leave (Tinto, 2006). Influential articles by Spady (1970) and Tinto (1975) opened discussions about associations between academic and social systems and student outcomes.

During the 1980s, the topic of student retention became prominent at national conferences. This improved the researcher's ability to access a large body of knowledge (BOK) being developed across the nation. This expanded knowledge led to new approaches to the study of retention (Berger & Lyon, 2005). By the 1990s, retention was firmly established as a critical issue within higher education. There was an increased focus on previous research, theoretical models and applying theory to practice. Retention studies at this time focused on two major categories: psychological persistence and social attainment. Researchers examined personality, motivation and intellectual factors to explain differences in persistence. Intelligence tests (IQ), scholastic aptitude tests (SAT) scores and the results of personality inventory tests were analyzed (St. John, 2000). However, during this period a number of researchers argued that studies which focused solely on scores often failed to control for external variables such as class level, type of institution, and background information. The social attainment camp debated that the majority of students who failed to successfully complete their college degree directly reflected the student's social, economic, or cultural background. This accounted for a growing number of studies focused on the influence various psychological and social forces had on student retention.

However, Tinto (1999) felt with increased attention to retention, institutions still did not take the issue seriously. He outlined a number of steps an institution should take to improve retention. He argued that institutions should move beyond the provision of "add-on" services and establish educational services that promotes retention for "all" students, not just some students who are considered at-risk based upon scores, personality tests, social, economic or cultural background. Although, Tinto acknowledged that the root of the retention issue depends on the student and the situations they face, he felt issues with retention was equally associated to the quality of the educational setting in which the student learns. In his seminal work "*Taking Student Retention Seriously"* (1999), the author recommended several conditions to improve retention at the institutional level. These conditions included support, advice, increased involvement, higher expectations and improvement of the learning environment. He emphasized the importance of applying these conditions for first year students.

In the early 2000s, retention literature stressed holistic approaches designed to support students. These strategies addressed both formal and informal student experiences inside and outside of the classroom (Demetriou & Schmitz-Sciborski, 2012). A wide range of studies revealed interactions students have with faculty, peers, and administrators directly influenced retention (Dringus, 2001; Habley & McClanahan, 2004; Kadar, 2001; Thayer, 2000) while a number of studies found learning style and motivation as contributing factors as to whether a student persisted or did not persist (Dringus & Terrell, 2000; Terrell, 2002).

With the rapid advancement of online learning environments (OLE), non-traditional students became the fastest growing population on college campuses throughout the United States. According to Brown (2002), this student population accounted for 50% of higher education enrollments in the early 2000s and has increased significantly to-date. Today, the non-traditional student is the new majority representing 75% of online enrollment for undergraduate and graduate programs (Council of Graduate Schools, 2010). Retention rates from the early 2000s until today consistently average between 10-20% lower in online learning programs at the graduate level (Carr, 2000; Council of Graduate Schools, 2008, 2012) compared to traditional on-campus programs.

Researchers continue to assess and examine student retention from different perspectives using a variety of techniques. The development of statistical models to examine and to forecast student academic progress using a combination of student factors is emerging as an innovative method to predict if a student is at-risk at the program or course level. Techniques for extracting knowledge from institutional data repositories allows researchers to build models which show promise for accurately reflecting student progress and outcomes in real time (Campbell, DeBlois,

& Oblinger, 2007). These methods have proven effective for improving retention rates at a number of universities.

2. Theoretical Perspectives of Retention

Although retention models have been influential in explaining student persistence in higher education, a number of models have been developed for traditional students enrolled in on-campus programs and were limited in explaining persistence of students studying at a distance. However, early works by Bean and Metzner (1985) Model of College Student Dropout and Moore's (1997) Transactional Distance Model were influential focusing on non-traditional students and addressing issues of distance. These seminal models predated modern online modalities and laid the foundation for future studies examining student retention specific to OLEs.

Bean and Metzner's (1985) Model of College Student Dropout (Figure 1) provides a framework of several broad categories representing a number of factors that are unique to non-traditional students. Categories include background, academic, environmental and social variables impacting academic performance (grades) and psychological outcomes (stress, satisfaction, goal commitments) that may influence a student's intent to leave and not successfully complete a course of study. To operationalize the model, Bean and Metzner (1987) conducted a study using a mixed methods approach. The authors surveyed over 600 part-time, undergraduate students enrolled at a commuter college. Questionnaire responses were derived from three sets of theories (Bean & Metzner, 1985; Locke & Bryan, 1968; Tinto, 1975) using regression analysis to analyze 26 variables potentially affecting retention. The study revealed grade point average (GPA), credit hours enrolled, age and race as having a significant impact on whether a commuter student successfully completes a course of study.

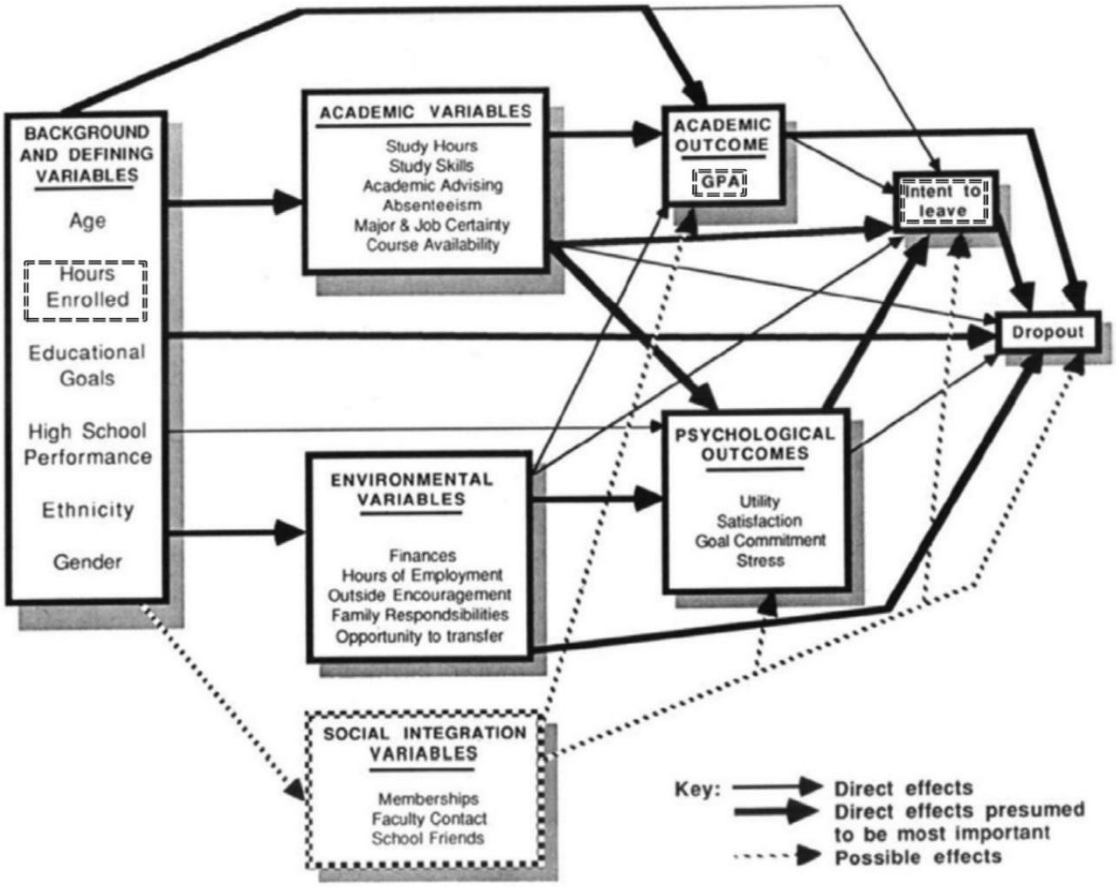


Figure 1. Bean and Metzner Model of College Student Dropout

Moore’s (1997) Transactional Distance Model (Figure 2) is based upon the psychological and communicative distance between instructors and students in an OLE. The degree of distance experienced by online learners can differ significantly with each individual student and with the environment in which they learn. According to Moore (1997), three factors significantly impact online learner’s experiences and outcomes: student autonomy, dialogue between instructor and student and structure of course design. The author contended that it is intuitive that physical distance between online learners and the institution or instructor can result in feelings of isolation and loss of motivation for some students.

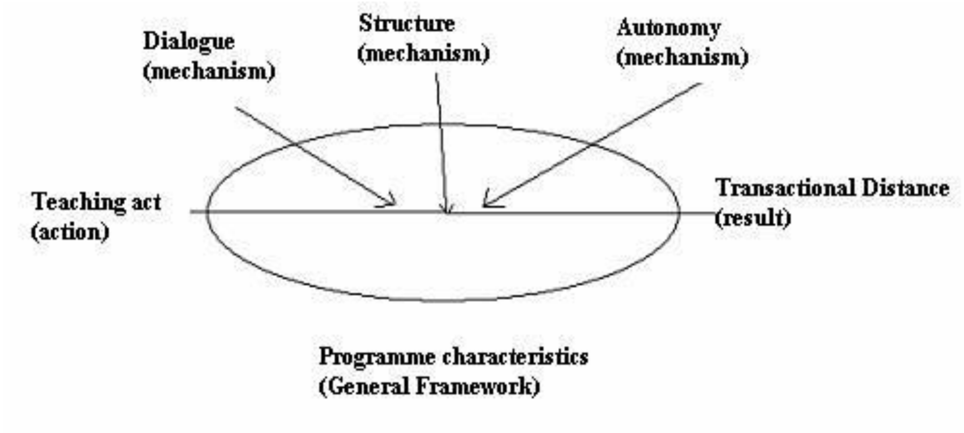


Figure 2. Moore Transactional Distance Model

In another seminal work, Rovai (2003) developed the Composite Persistence Model (CPM) to predict persistence of non-traditional students studying at a distance (Figure 3) by synthesizing retention models developed by Bean and Metzner (1985) and Tinto (1975, 1993). Rovai proposed that although a number of theoretical models had paradigmatic stature, retention models were largely based upon psychological attributes which minimally examined factors based upon student fit, attributes prior to admission and external and internal factors unique to online learners. The CPM included age, ethnicity, gender, academic performance, literacy, written performance and interaction skills. Additional factors such as finances, employment status, family responsibility and life crises such as sickness and divorce were included in the model. Rovai's model has also been influential in directing teaching strategies and promoting programs to improve retention (Gazza & Hunker, 2014).

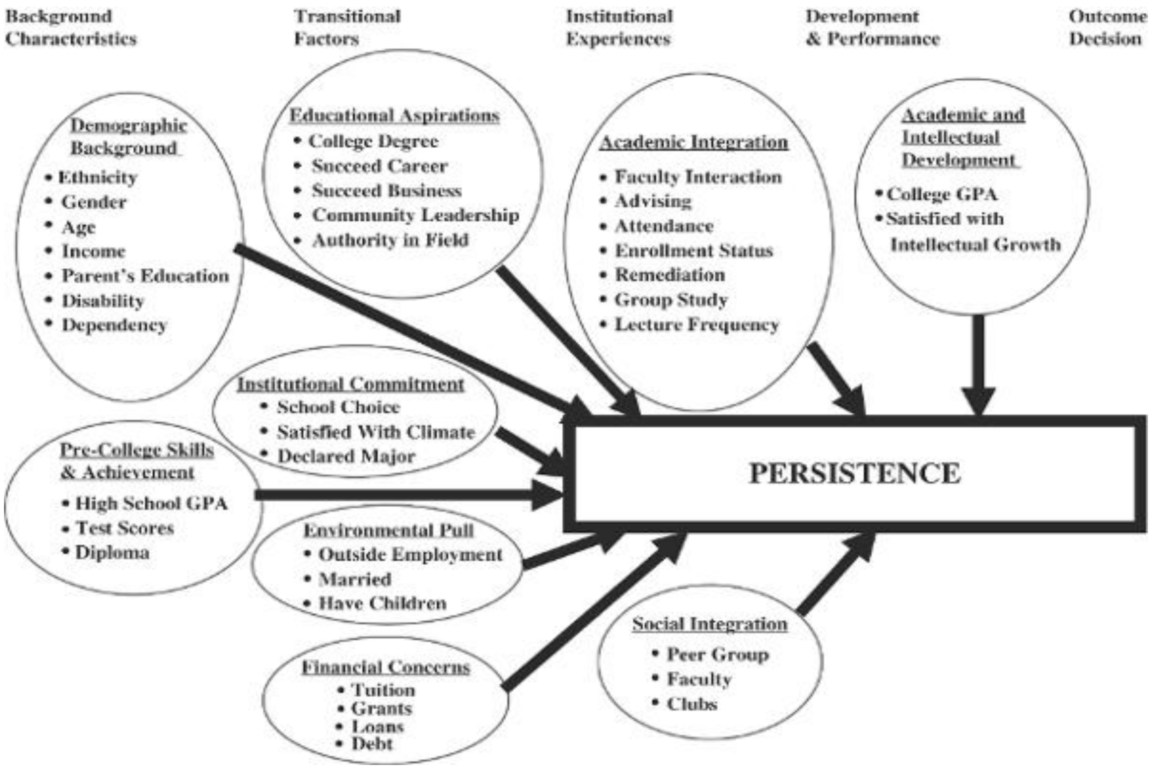


Figure 3. Rovai Composite Persistence Model as illustrated by Freeman (2003)

Recent studies have provided empirical evidence supporting the original design of Rovai's CPM model. Perry, Boman, Care, Edwards, and Park (2008) qualitative study investigated students self-identified reasons for not successfully completing online graduate programs in nursing and health using the CPM model as a framework for analysis. The major reasons for not persisting fell under two categories: external and internal. External factors included finances, hours of employment and family commitments. Internal factors included academic integration and institutional factors such as program of study. The authors noted that for this particular study there was no evidence that the non-traditional student population did not persist because of a perceived lack of social integration or absence of a community of learning.

In a more recent study, Lee, Choi, and Kim (2013) examined the differences between students who persist in an online course and students who do not successfully complete the course based upon background, transitional, institutional and performance factors unique to Rovai's model. Results of the study revealed that entry and background characteristics were significant in identifying students who may be at-risk. The authors suggest assessment of these factors at the beginning of a course is critical in order for instructors and administrators to provide the necessary support early in the semester.

To date, there is no consensus as to which theory and associated factors are most relevant to fully explain retention of non-traditional students studying at a distance. The evolution of student retention theory and practice has expanded from a programmatic approach and has evolved integrating models that position each student in a position for success (Habley & McClanahan, 2004). With existing theories on student retention firmly established, previous research has identified factors that can be associated with a student not completing an online program as well as a student not successfully completing an individual course. Factors identified in Rovai's CPM model provides a basic framework for this study.

3. Non-traditional Student Attributes

Although there is no precise definition of a non-traditional student, the National Center for Educational Statistics (NCES, 2013) suggests that common characteristics of the non-traditional student is based upon the following elements: age, part-time status, delayed enrollment, full-time employment, financial independence, dependents and completion of high school with a general education diploma (GED). In a recent NCES publication by Aud and Wilkinson-Flicker, *The Condition of Education* (2013), fall enrollment for post-secondary education in 2011 accounted for 71% of full-time students and 78% for part-time students who were at least 25 years or older. Changing work demands, financial challenges and the desire for professional advancement fuels a student's enrollment in higher education (Kelly and Strawn, 2011).

By analyzing combinations of student data, you can identify not only sets of factors that impede desired outcomes for students enrolled in an online course, you can also identify positive factors that contribute to those outcomes (Fusch, 2011). Beyond skills such as grades and completed assignments (chapters), this study will examine student characteristics and internal factors which are based on Rovai's CPM model (2003). Student characteristics for this study will include age, gender, level of education, elapsed time between high school graduation and registration. Internal factors will consist of frequency of events, number of days student interacted with the course, number of video events, number of chapters (assignments) successfully completed and number of posts in the discussion forum at the end of the course.

Justification and selection of each variable or the combination of variables is based upon multiple theories, empirical evidence and domain knowledge. Construct consideration is relevant in the theoretical development of the PMS model.

3.1. Student Characteristics

3.1.1. Age

Age is often included as a control variable in research examining retention of non-traditional students enrolled in online courses. Previous research has revealed that as the average age of the college student increases, the risk for not successfully completing a course or a program of study rises (Bean & Metzner, 1985; Horn, 1998; Stratton, O'Toole & Wetzel, 2007). Many older students have more responsibilities outside of school such as work and family obligations. As a result, some students will not persist (Bean & Metzner, 1985). It was also revealed in a survey of the literature conducted by Dobbs, Waid and del Carmen (2009) that there is a significant difference in perceptions for students enrolled in online programs compared to traditional students attending classes on-campus indicating that age is also a risk factor for not successfully completing online courses.

3.1.2. Gender

Females, on average, outnumber males in post-secondary education and academic performance (Severiens & ten Dam, 2012). According to a study conducted by Jameson and Fusco (2014), this has been the trend from the 1990s onward in the majority of western countries. Nationally, at post-secondary institutions, completion rates were higher for female students (Snyder & Dillow, 2012). However, data reveals that graduation rates are significantly lower for females enrolled in science, technology, engineering and mathematics (STEM) courses (NSF, 2006). Significant differences arise when examining groups who initially declare a STEM major. Males (31.8%) in a sample declared a STEM major compared to 14.3% of female students. The National Science Board (2010) Science Indicators report reveals females who begin college as STEM majors have a lower probability of receiving a degree in a STEM field. Historically, females are the least likely to persist toward a degree in one of these fields.

Although, various contributing factors have been examined, gender disparities in STEM courses still exist due to perceived marginalization or bias that women experience in co-educational settings with peers and professors (Rosenthal, London, Levy & Lobel, (2011). In the past ten years, according to the National Science Foundation (2014), males continue to earn more bachelor’s degrees in engineering, computer science and physics. These differences are largely, but not entirely, due to higher enrollment of males.

3.1.3. Delayed Enrollment

There is growing interest in the research field as to how elapsed time between high school graduation and post-secondary enrollment influences whether a student will persist in an online program. The typical first-generation student is more likely to delay entry, begin at a two-year institution, attend part time, and attend discontinuously (Chen & Carroll, 2005; Ishatani, 2006; Tinto, 2012). Life transitions, including employment and family obligations make a unique contribution to explaining delayed enrollment (Wood, Kurtz-Costes, & Copping, 2011). Studies conducted by Grubb (1997) and Horn and Carroll (1996) reveal how combined factors of delayed enrollment, employment hours and family obligations had negative effects on the probability of a student completing their degree. In addition, Bozick and DeLuca (2005) found that for every month of post-secondary enrollment delay, students had a lower probability of successfully completing a program. The results of the study also revealed for every one year of delayed enrollment, students had a 48% lower odds-ratio for graduating from a program. Lastly, findings from research reveals length of delayed enrollment results in lower levels of academic readiness and integration decreasing the likelihood of persisting and attaining a degree (Calcagno, Bailey, Jenkins, Kienzl, & Leinbach, 2008; Pascarella & Terenzini, 2005; Tinto, 1993).

3.3. Internal Factors

3.3.1. Engagement in LMS

The growing use of learning management systems (LMS) in OLEs provides researchers access to student activity through logged data automatically stored in the LMS system. During the past decade, educational researchers (Agudo-Peregrina, Iglesias-Pradas, Conde-Gonzalez & Garcia, 2014; Hung & Zhang, 2008; Black, Dawson & Priem, 2008; Terrell, Snyder & Dringus,

2009) investigating attrition and retention have employed data mining techniques to gain insight about student performance from online activities extracted from LMS data repositories. However, according to Hu, Lo, and Shih (2014), the number of studies examining these time-dependent variables is limited. As a result, the authors selected an online course and measured how time-dependent variables impact final outcomes. Variables for this predictive study included: course login count, course login time average, and course login date/time. Course login time/date and course login time average ranked 1st and 3rd, from thirteen variables examined, as significantly influencing final grades for the course.

In a similar study, Coldwell, Craig, Paterson, and Mustard (2008) examined the relationship between early participation and student performance. The authors found a relationship exists between student engagement and academic performance measured by final grades. The results also suggest that by tracking logging data early in the course, this data can be used as an early indicator for identifying students who are at-risk early in the semester.

3.3.2. Frequency of Engagement

Moore (1989) identified three types of academic engagement in the OLE: learner-content, learner-instructor, and learner-learner. These interactions support both instructional and social goals by establishing collaboration among class members and the instructor. Researchers recognize that community building serves two purposes. It provides a sense of togetherness and also helps to keep students engaged in the class (Brown, 2002). According to Rovai (2002) and Terrell, et. al., (2009), faculty and students must continually communicate with each other to build a strong sense of community. If a student feels they are not accepted and lack a sense of safety and trust with class members and faculty, they will not feel connected to the learning environment. However, systematically quantifying frequency of posts in the LMS is complex due to a combination of factors such as length of time logged in, quality of posts and individual time constraints influencing frequency of posts. In a study conducted by Macfayden and Dawson (2010), the authors utilized multiple linear regression analysis to determine which factors influence academic outcomes. Fifteen communication variables from LMS data usage logs were analyzed in order to generate a best-fit model to identify students who are at-risk for not completing a course. An 81% prediction accuracy was obtained identifying students who were at-risk based upon final grade. Total number of discussion posts and successful completion of assignments were ranked as key variables supporting the predictive power of the model.

In a similar study conducted by Kupczynski, Gibson, Ice, Richardson, and Challoo (2011), the author's goal was to examine if there was a relationship between frequency of participation and student achievement as measured by the final grade in a course. While the impact on achievement resulted in a 10.1% variance, the authors concluded that participants in the study who posted with greater frequency achieved a higher level of success in the course as measured by final grades.

4. Predictive Modeling

Predictive modeling is a commonly used statistical technique by which a model is developed to best predict the probability of an outcome (Geisser, 1993). Predictive models are utilized directly to estimate future behaviors given a defined set of attributes (input) or indirectly based upon decision rules (Steyerberg, Vickers, Cook, Gerds, Gonen, Obuchowski, & Kattan, 2010). Historical and current data is collected and analyzed to formulate a model. Development of the model is reiterative where the model is often revised based upon the accuracy of the results and the availability of new data.

Early modeling research utilizing statistical techniques to analyze multiple-variables from student data repositories can be traced to studies conducted by Aitken (1982) and Pascarella and Terenzini (1980). In the early 1980s Aitken (1982) and Pascarella and Terenzini (1980) proposed developing a multi-equation model to operationalize the underlying structural relationships that determines academic outcomes. While a complete structural model was not detailed in either study, the authors suggested a need to combine sets of variables based upon theory and from findings presented in seminal works examining single attributes.

Expanding on these works, Murtaugh, Burns, and Schuster (1999) conducted a longitudinal study utilizing the Cox proportional hazards regression model (Cox, 1972) to predict a student's (*n* = 8,867) probability of leaving school based upon a combination of ten demographic and academic variables. Independent associations of race/ethnicity, class level and age of student (25+) were found to influence lower retention rates significantly. In a similar study, McDaniel and Graham (1999), developed a prediction model using stepwise regression which involved starting with one variable and testing the addition of each variable for accuracy. The model included 25 external and internal factors to predict the retention status of 1,949 freshmen students who entered the institution from 1990 to 1995. Results of the study revealed returning students had significantly higher ACT scores, high-school GPA, and cumulative GPA compared to students who did not return for the second year of college.

Predictive modeling studies focusing on non-traditional students enrolled in online courses began to emerge in the early 2000s with the popularity of online programs. Minaei-Bidgoli, Kortemeyer, and Punch (2004) developed a system that could routinely collect vast quantities of information extracted from logged data within campus systems. The authors developed a genetic algorithm (GA) to classify variables in order of predictive accuracy. The GA demonstrated a significant improvement between 10 to 12% in identifying students who are at-risk as compared to modeling techniques with non-GA classifiers. Similar studies emerged using discriminant function analysis (Martinez, 2001), binary logistic regression (Woodman, 2001), Markov student-flow analysis (Herrera, 2006), regression analysis (Onwuegbuzie, Witcher, Collins, Filer, Wiedmaier, & Moore, 2007) and linear regression (Ayán & Garcia, 2008).

A variety of analytical approaches have been employed to improve prediction in recent years. Anaya and Boticario (2011), Baker and Yacef (2009), and Lopez, Luna, Romero, and Ventura (2012) demonstrates accuracy of outcomes using classification and clustering approaches to identify students who are at-risk. Delen (2010) developed an analytical model using ensembles to accurately predict if students would persist in their freshmen year in college. Macfayden and Dawson (2010) demonstrated successful correlation of 15 variables with final student grades using regression modeling.

Despite the growing number of studies focused on the development of models to predict academic performance, there is diversity among the research community on which analytical approach should be utilized and which combination of factors influence student outcomes. As a result, a few researchers have undertaken the task of comparing multiple techniques to determine which approach is the most appropriate to predict student performance. Akçapınar, Coşgun and Altun (2013) compared random forest decision tree, support vector machines (SVM), naïve Bayes and boosted classification tree algorithms to predict final grades. According to their findings, SVM outperformed other methods. In a similar study, Watkins (2013) compared approximate nearest neighbor (ANN), SVM and CHAID decision tree. Comparison results revealed that SVM also provided greater accuracy as compared to ANN and CHAID. Huang and

Fang (2013) found similar results comparing multiple linear regression (MLR), multilayer perception network (MLP), radial basis function (RBF) and SVM models.

Although, SVM produces the highest overall accuracy, studies utilizing machine learning techniques reveal a number of limitations. How researchers perform hypothesis testing, such as K-Fold Cross Validation (K-CV) and handle multicollinearity, high error rates, outliers, and missing values are not detailed in a number of studies (Freckleton, 2011; Fürnkranz,, Gamberber & Lavrač, 2012). Secondly, the training and testing phases performed by specialized third-party software requires a lengthy time to develop and test. As a result, the prediction process is impeded impacting the models ability to predict performance in real-time in order to provide educational interventions (Guo & Paquet, 2013). Finally, predictive models utilizing machine learning techniques are not well supported by relational database management systems (RDBMS) despite their growing prevalence and importance in studies investigating student retention (Akdere, Cetintemel, Riondato, Upfal, and Zdonik, 2012). As a result, recent works on custom integration within educational systems are emerging to improve predictive performance and usability utilizing declarative languages to build simplistic analytical models embedded in the RDBMS.

The earliest integrated predictive modeling system can be traced to Purdue University through the implementation of Course Signals (CS). CS, an early warning system, was pioneered by Campbell, Deblois, and Oblinger (2007) and implemented by Arnold and Pistilli (2012) in the Spring of 2009. The premise behind the system was to develop an automated tool that could be accessed by instructor's to identify online students who were at-risk while a course is in progress. Since the initiation of the pilot project, retention rates improved significantly. In a similar project, Lauria, Baron, Devireddy, Sundararaju, and Jayaprakash (2012) developed the open academic analytics initiative (OAAI) system. The author's goal was to develop an open-source, automated predictive system that was accessible across a number of state-wide university systems. Results from the CS and OAAI projects report between 82% to 90% accuracy rates in predicting student outcomes. Limitations to the pilot projects include reliance on application program interfaces (API) extensions to transfer and transform data to facilitate the execution of algorithms for analysis potentially impacting performance and usability. Further drawbacks include the systems inability to categorize and analyze data by assigning weights to single factors or combinations of factors according to risk categorization association.

Model-based predictive methods is emerging as an alternative to current systems. A handful of systems show promise in extending a RDBMS to facilitate efficient real-time processes that are non-reliant on complex machine learning algorithms. Akdere, Cetintemel, Riondato, Upfal, and Zdonik (2012) developed a predictive database management system (PDBMS) prototype designed with two interfaces. The first interface consists of access to a Web-based tool targeted towards advanced users who want exert a hands-on control of the PDBMS and its associated operations. This approach provided an easy and effective way of utilizing and maintaining pre-tested and optimized logic within the RDBMS framework utilizing SQL query language. The second interface access method provides experts the ability to maintain SQL functions consisting of extraction procedures, variable assignments, analysis processes and hypothesis testing. In a similar pilot project, Graf, Ives, Rahman, and Ferri (2011) developed the Academic Analytics Tool (AAT) designed to allow instructors to perform simple to complex analytical queries on student data using a Web-based tool. The open-source tool was designed to run independently across a variety of educational systems and operate with Moodle, Sakai, and Desire2Learn LMS systems. More recent works include Guruler and Istanbullu

(2014) Web-based predictive software system using knowledge discover in databases (KDD) methodologies. Arnold and Campbell (2013) continue to work on Course Signals (CS) system with improvements to automate collection, analysis and categorization of data.

This study proposes a Web-based predictive modeling system utilizing RDBMS, SQL, and inductive logic programming (ILP) to efficiently identify students at-risk. A gap exists in educational research examining these methods in combination. However, similar systems reliant on these methods have been researched in the medical field to predict patient outcomes (Peissig, Santos-Costa, Caldwell, Rottscheit, Berg, Mendonca, & Page, 2014; Qiu, Shimada, Hiraoka, Maeshiro, Ching, Aoki-Kinoshita, & Furuta, 2014).

Chapter 3

Methodology

The development of the predictive modeling system (PMS) is guided by the systems development life cycle (SDLC) waterfall model approach (Figure 4) originally developed for information technology projects by Royce (1970). The waterfall model consists of the following phases: requirements, design, coding, testing and integration.

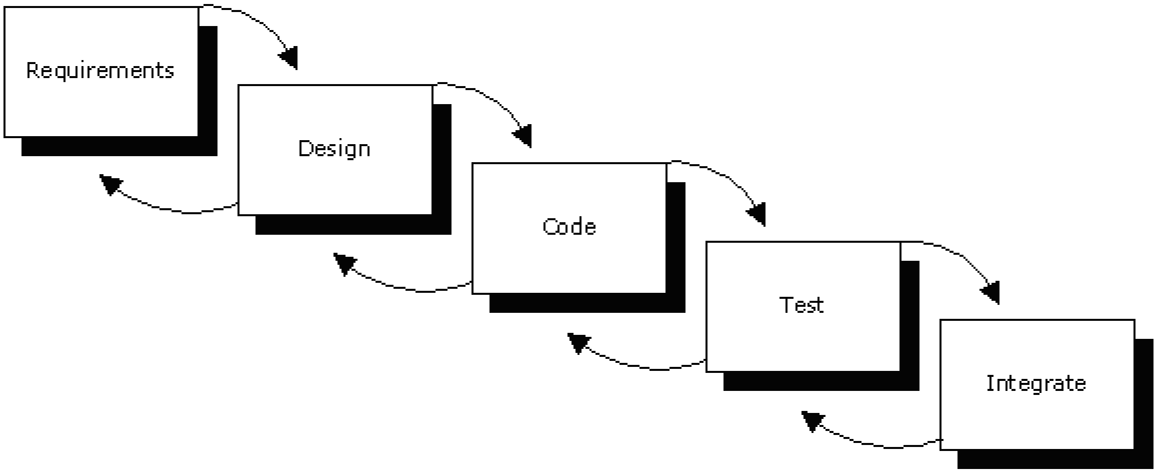


Figure 4. SDLC Waterfall Model

This method follows a structured approach having a logical flow of development activities. The *requirements* phase discusses the data source and the sample population. It also outlines the system architecture (Figure 5) defining the hardware and software components required to build the PMS model. The *design* phase illustrates the PMS model which consists of two Web sites including displayed results. The *coding* phase discusses the development of the initial Web site (Figure 6) written in HTML5, the user selection screen (Figure 7) written in PHP in conjunction with embedded PDOs (API /CGI extensions) and SQL statements for data extraction, analysis and reporting. The *testing* phase defines analysis methods employed, variable hierarchy, student population and validation techniques utilized. The *integration / implementation* phase includes an objective description of the findings describing the statistical techniques applied to the data, interpretation of results, conclusions that were drawn including implications and recommendations. During this phase results are discussed in terms of its relation with results obtained in previous research.

1. Requirements

1.1. Data source and sample population

Data for this study was extracted from Harvard and Massachusetts Institute of Technology (MIT) universities massive online open course (MOOC) “edX” dataset. The publicly available dataset contains student records from seventeen courses. The collaborative effort between Harvard and MIT was jointly founded to increase learning opportunities for students worldwide and to advance educational research. Harvard offered six courses while MIT offered eleven courses on the same platform (edX) for the academic year ranging from Fall 2012 to Summer 2013. Subjects included biology, chemistry, computer science, electronics and engineering courses. The MIT/6.002x/Fall_2012 Circuits and Electronics selected for this study drew 40,811 registrants.

Dataset de-identification and compilation was processed by Ho, Reich, Nesterko, Seaton, Mullaney, Waldo, and Chuang (2014) following strict Federal government guidelines in order to protect student privacy put forth by the Family Educational Rights and Privacy Act (FERPA) (20 U.S.C. § 1232g; 34 CFR Part 99), (D.O.E., 2014). The de-identified dataset (AY2013) was released for public accessibility in June, 2014 (HarvardX - MITx Dataverse Network, 2014).

Data preparation

Analysis of the initial dataset (*n* = 6,566) for the selected MIT/6.002x/Fall_2012 Circuits and Electronics course exposed missing values in the following fields: year of birth, level of education, last activity date and gender. A final grade of “0” existed for 4,385 records. Harvard / MIT data definitions did not clarify if “0” denoted an incomplete grade or if the course was being audited. For this specific course, it was determined that missing values or a final grade of “0” for this particular dataset would compromise the analysis process for identifying if a student was at-risk for not successfully completing a course.

Students who registered for the Circuits and Electronics course represented 150 countries world-wide. Definitions of retention of non-traditional students in post-secondary education were not applicable to students who registered from countries other than the United States. Researchers at M.I.T. and Harvard determined the country of residency by capturing the internet protocol (IP) address when students registered for the course. Resulting in a sample size of 1,804 records. Of the 1,804 records in the sample population, 67% did not participate in the course after registration. This was indicated by a “0” in the grade field. Of the remaining 33% (568), 24% of the population was eliminated due to missing data fields such as: year of birth, gender, event frequency, video views and number of days active. The final dataset (*n*=175) consisted of 9% of the original dataset.

As a result, the final dataset was reduced to students (*n* = 175) who were considered non-traditional based upon age, who reported being enrolled or having completed a Bachelor of Science degree who resided in the United States.

1.2.Analysis of data

Four types of analytical approaches were conducted for this study using SPSS version 22 to analyze results. Initially two analytical methods were performed on the data. Descriptive analysis was concerned with the investigation of individual factors in regards to its effect on student's actual grade and to the predicted outcome by reporting frequency and mean and by using cross tabulation to examine the results (totals) for independent variables for the entire student population in order to find relationships between variables. These variables included: age, chapters completed, delayed enrollment, event frequency (clicks), gender, days student interacted with the course, initial start date (as compared to course start date) and video events (clicks) including . If a student's grade is 55% or above the student's certification field was automatically updated by M.I.T. with a "1" in the original data set. Additionally, feature selection was conducted as a measure of statistical dependence (importance) factors have on student's final grade.

Multiple regression analysis was performed during training and testing (80:20) to predict the value of Y (predicted outcome) for the values of X_1, X_2, …, X_k. (given by: $\boldsymbol{Y = b_0 + b_1 X_1 + b_2 X_2 + \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots + b_k X_{k)}}$). The appropriateness of the multiple regression model was tested using ANOVA f-test to determine how well the data fits each model (formula). This process was reiterative in order to determine what formula or formulas during training is the "best fit" for the final test phase. In conjunction with multiple regression analysis, a paired sample t-test was utilized to compare the actual student grade to the predicted outcome throughout each training and testing phase.

1.4. System overview

The system diagram (Figure 5) is a graphical representation of the predictive modeling system (PMS) requirements. Web components will conform to current standards set forth by Web content accessibility guideline technical standards (Caldwell, Cooper, Reid & Vanderheiden, 2008) and the World Wide Web Consortium (W3C) technical specifications.

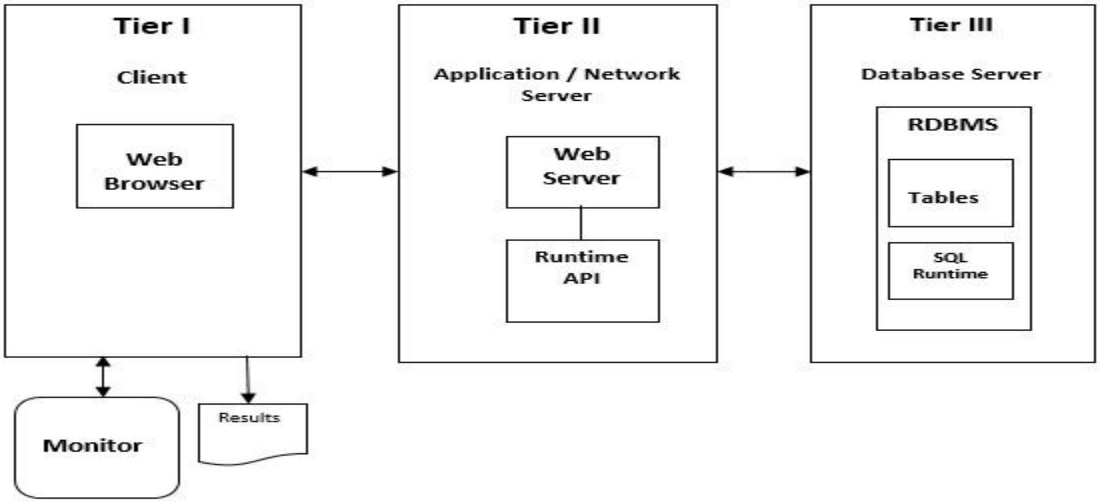


Figure 5. PMS Architecture

Tier I consists of a single client or multiple clients. The client refers to the user interface (PMS) and Web browser which will run locally on a workstation. The Web browser initiates communication with the application server (Tier II) which receives content from the client. The primary function of the application server is to store (application files), process (PDOs and SQL commands) information based upon content received from the client using the hypertext transfer protocol (HTTP) and delivering this information in the form of requests between the client and database server (Tier III). The database server is the relational database management system (RDBMS) consisting of student tables, scripts, and SQL commands. Analysis of data is processed within the RDBMS. Results of analysis (output) is returned to the client (user) and displayed on the monitor with an option to print utilizing hypertext preprocessor language (PHP) echo and print statements. According to Connolly and Begg (2010), a three-tiered design has a number of advantages which includes:

- The need for less expensive hardware because the client is ‘thin’.
- Application maintenance is centralized transferring logic between client and database.
- Ease of replacement of individual tiers resulting in compatibility with other systems.
- Load balancing logic between application server and database server is efficient.

2. Design

The PMS Web interface is designed to provide instructors with the flexibility to select courses, students and factors. When initially accessing the user interface, instructors log-in to the system using instructor identification and password (Figure 6). For the purpose of this study, three instructor identifications and passwords were setup for initial testing. The username and password is validated on the client side using HTML5 code. If an incorrect username or password is entered the instructor receives a message indicating either one or both fields are invalid.

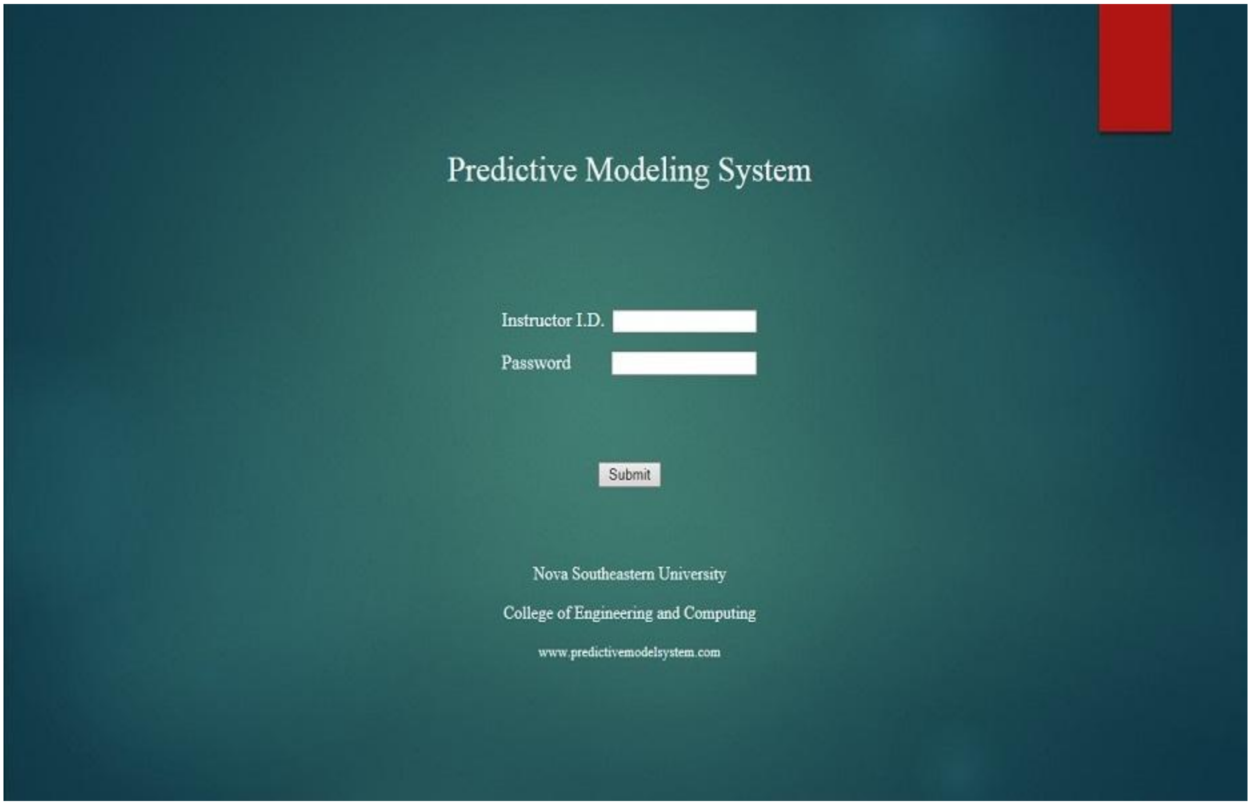
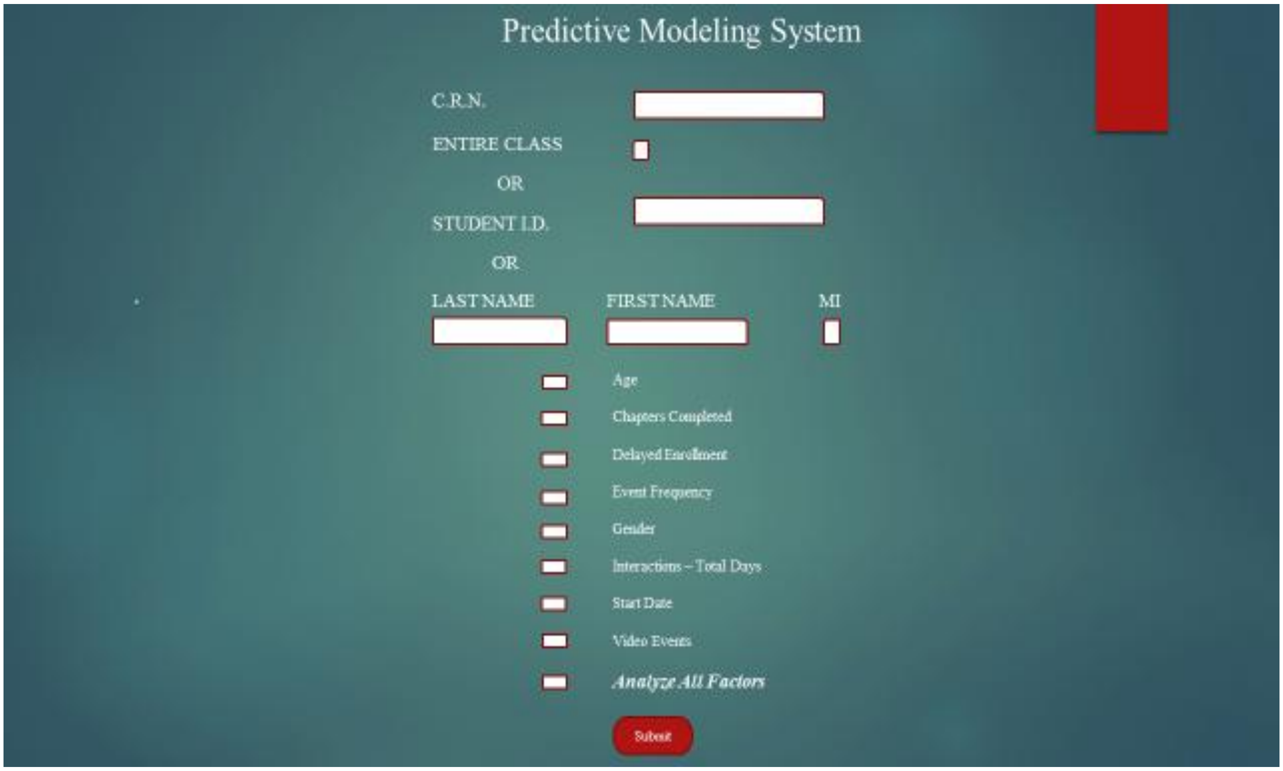

Predictive Modeling System
Instructor I.D.
Password
Submit
Nova Southeastern University
College of Engineering and Computing
www.predictivemodelsystem.com

Figure 6. PMS Log-in

The second access screen (Figure 7) prompts instructors to enter the course registration number (CRN). Selecting the entire class or selecting an individual student by I.D. or name is optional. The interface provides instructors with two options. The first option is to select all factors (analyze all factors). This is the default setting which runs an embedded formula derived from training and testing data. The instructor will also have the option to select a specific factor or a combination of factors that is unique to the specific class or individual student. The ability to select a subset of variables for specific analysis is an additional feature unique to the PMS model. Results from running the PMS model is displayed on the screen (Figure 8) with an option to print. The “at-risk” field with a “0” denotes that the student or students are potentially at-risk for not completing the course.



Predictive Modeling System

C.R.N.

ENTIRE CLASS □

OR

STUDENT I.D.

OR

LAST NAME

FIRST NAME

MI

□ Age

□ Chapters Completed

□ Delayed Enrollment

□ Event Frequency

□ Gender

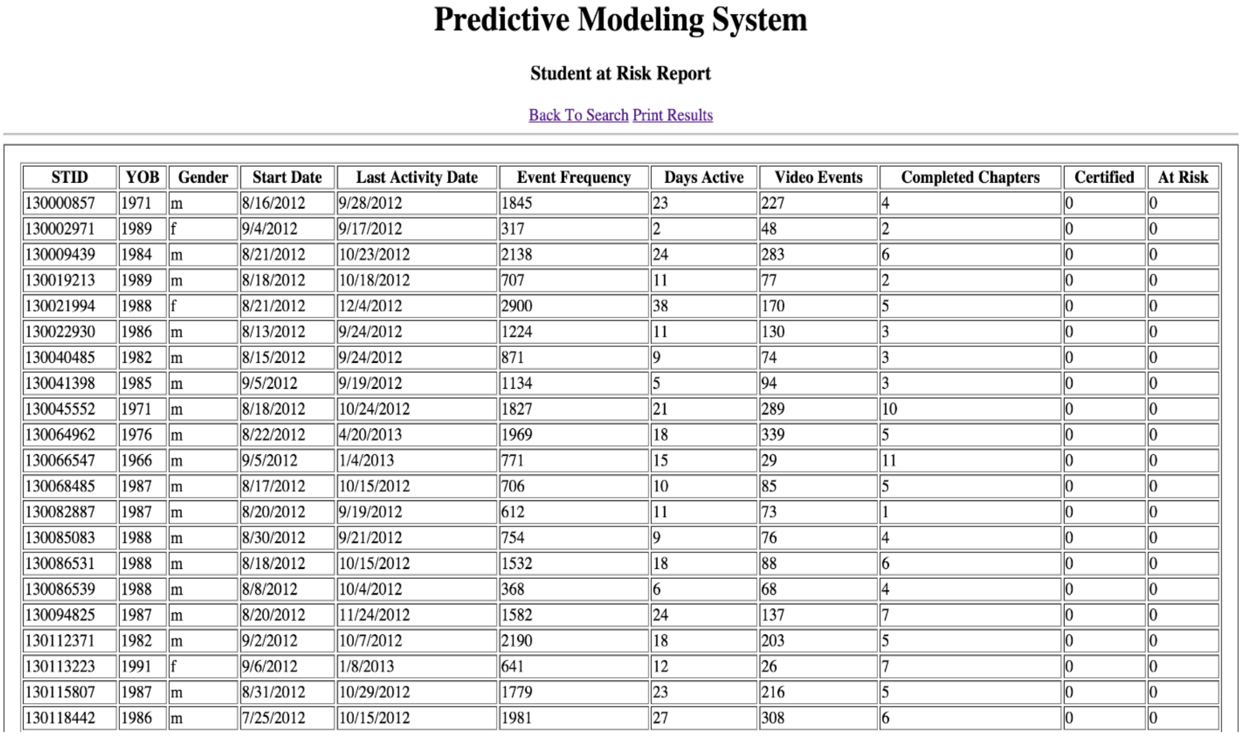
□ Interactions – Total Days

□ Start Date

□ Video Events

□ *Analyze All Factors*

Figure 7. PMS Selection Screen



Predictive Modeling System

Student at Risk Report

STID	YOB	Gender	Start Date	Last Activity Date	Event Frequency	Days Active	Video Events	Completed Chapters	Certified	At Risk
130000857	1971	m	8/16/2012	9/28/2012	1845	23	227	4	0	0
130002971	1989	f	9/4/2012	9/17/2012	317	2	48	2	0	0
130009439	1984	m	8/21/2012	10/23/2012	2138	24	283	6	0	0
130019213	1989	m	8/18/2012	10/18/2012	707	11	77	2	0	0
130021994	1988	f	8/21/2012	12/4/2012	2900	38	170	5	0	0
130022930	1986	m	8/13/2012	9/24/2012	1224	11	130	3	0	0
130040485	1982	m	8/15/2012	9/24/2012	871	9	74	3	0	0
130041398	1985	m	9/5/2012	9/19/2012	1134	5	94	3	0	0
130045552	1971	m	8/18/2012	10/24/2012	1827	21	289	10	0	0
130064962	1976	m	8/22/2012	4/20/2013	1969	18	339	5	0	0
130066547	1966	m	9/5/2012	1/4/2013	771	15	29	11	0	0
130068485	1987	m	8/17/2012	10/15/2012	706	10	85	5	0	0
130082887	1987	m	8/20/2012	9/19/2012	612	11	73	1	0	0
130085083	1988	m	8/30/2012	9/21/2012	754	9	76	4	0	0
130086531	1988	m	8/18/2012	10/15/2012	1532	18	88	6	0	0
130086539	1988	m	8/8/2012	10/4/2012	368	6	68	4	0	0
130094825	1987	m	8/20/2012	11/24/2012	1582	24	137	7	0	0
130112371	1982	m	9/2/2012	10/7/2012	2190	18	203	5	0	0
130113223	1991	f	9/6/2012	1/8/2013	641	12	26	7	0	0
130115807	1987	m	8/31/2012	10/29/2012	1779	23	216	5	0	0
130118442	1986	m	7/25/2012	10/15/2012	1981	27	308	6	0	0

Figure 8. PMS Displayed Output

3. Coding

When the analyze request is submitted, query tasks include extraction and analysis of data based upon user selection. Analysis formulas have a hierarchal order where factors are ranked by the highest to the lowest predictive power based upon extant literature and domain knowledge. The analysis framework subsumes first-order logic based upon the principles of statistical relational learning (SRL) and associated principles of inductive logic programming (ILP). This machine learning (ML) approach utilizes SQL declarative language integrating basic concepts from ILP through constraint logic programming and inductive reasoning resulting in a flexible environment for predicting outcomes. This approach is motivated by the view of data mining (DM) as a querying process originally proposed by Imielinkski and Mannila (1996) and demonstrated by Fu (2011), Kantardzic (2011) and Trasarti, Giannotti, Nanni, Pedreschi, and Renso, C. (2012). The following formula is embedded within the PMS selection screen utilizing

HTML5 (client side) and PHP / PDO (server side (Web and SQL server)) code. The code is activated based upon selection of the entire class for initial testing of the predictive modeling system (PMS).

Formula

```
          /* calculate whether the student or students is at risk
           */
function formula2($data) {
   $daysActive  = $data['daysact'];
   $startDate   = strtotime($data['regdate']);
   $lastActDate = strtotime($data['laactdate']);
   $compareDate = strtotime('2012-10-15 00:00:00');
   $yob         = $data['yob'];
      if($daysActive <= 27 /*&& $lastActDate < $compareDate*/ && $yob >= 1982) {
      return true;
   } else {
      return false;
   }
 };
```

In this formula, the “risk status” field created in the student record table will be updated as “at-risk” with a “0” if: (condition 1) days active is less than or equal to 27, (condition 2) if last activity date is less than October 15, 2012 and (condition 3) if year of birth is greater than or equal to 1982. If all conditions are “not true” the student will not be flagged at-risk and the at-risk field in the output file will equal a “1”.

According to de Raedt (1998), this approach puts inductive logic programming into a new perspective. SRL extends the search space to include a richer set of features, including many which are not Boolean, where the model and search selection are integrated into a single process allowing information criteria, native to statistical modeling, by making selection decisions in a step-wise manner.

Risk values outlined in Table 1 are derived from seminal works investigating retention of non-traditional students enrolled in online programs. In a recent study conducted by Ho, Seaton, Reich, Nesterko, Mullaney, Waldo and Chuang (2014), data was collected from four online courses including the Fall 2012 MIT6.002 dataset under investigation. Results revealed a significant association between grades and factors listed in Table 4. Ho, et. al., found a few courses lacked vital information such as updated grades and certification fields. However, the facilitators for MIT6.002x consistently updated the certification field with a "1" if the student passed with a final grade of 55% or a "0" if the student did not successfully pass the course.

Rank	Field Name	Values	Risk Values
1	Chapters Completed	1 - 18	< 14
2	Event Frequency (Key – Strokes)	31 – 2,218	< 3,120
3	Interaction (Total Days Active)	1-151	< 27
4	Video Events (Clicks)	1-4,289	< 373
5	Gender	M, F	M
6	Delayed Enrollment (Years)	0-38	< 19
7	Age	20 – 61	< = 1982
9	Months Engaged	(1-360)	<=90
10	*Final Grade* (55% + = Certification)	(.01 – 1.00)	< 55 %

Table 1. Factor Rank

Formulas are represented in conjunctive queries with an equation for each possible state or states. Independencies can also be viewed as compactly representing a factorization of joint probabilities based upon values guided by these formulas. An illustration of conjunctive formulas based upon threshold risk values defined in Table 2 is demonstrated in Table 1.

Formula	Student ID	Student Data	Predicted Risk Factor
1	130082887	Day Active=11; Last Active Day=20120919; YOB=1987	0
1	130300185	Day Active=17; Last Active Day=20121014; YOB=1983	0
1	130459311	Day Active=56; Last Active Day=20130209; YOB=1962	1
1	130379779	Day Active=27; Last Active Day=20130703; YOB=1959	1

Table 2. Conjunctive Formula Initial PMS 7Model Test

4. Test

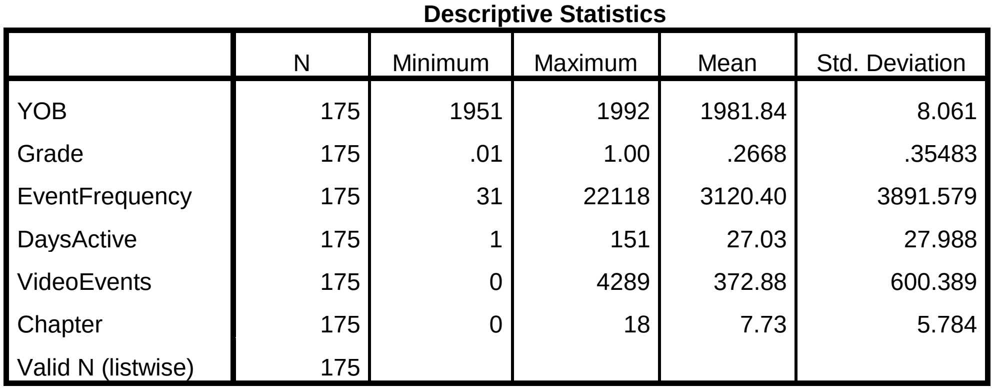
In the test phase a heuristic evaluation of the working model is conducted by experts in the field of educational technology. This phase will act as an anchor to evaluate PMS performance and to make final modifications to hardware, software and associated formulas. The overarching questions during this phase includes: Is the PMS usable? In order for the PMS to be considered usable, the model should be efficient, effective, useful, and accessible (Rubin & Chisnell, 2008). A subset of research questions will include: (a) Does the PMS allow the user to easily access the system? (b) Does the PMS process in a way that a user expects? (c) Can a user operate the PMS to a defined level of competence? (d) Does the PMS system produce accurate results identifying students who are at-risk?

To build the model, training and testing involved using a 60:40 ratio split. Forty percent of the data (70 records) are set aside for testing (validation). Training entails running four formulas over 60% (105 records) to observe performance (accuracy of prediction). Accuracy of prediction during each phase is measured by comparing the student certification field to the predicted outcome field. A “0” in in the certificate field indicates the student did not receive certification and did not successfully pass the course. If a student is predicted to be at-risk a “0” is output in the at-risk field. Accuracy of prediction and validation of the PMS model was measured using a paired sample t-test, frequency distribution, cross-tabulation, and regression correlation utilizing SPSS version 22.

Chapter 4

Results

Retention for students enrolled in M.I.T. Circuits and Electronics course in the Fall 2012 semester should be considered in the context of learner intent which differs from the non-traditional student enrolled in a credited online course. When viewed in the appropriate context, the low retention rate (22.3%) for this massive open online course (MOOC) is considered reasonable (Koller, Ng, Do & Chen, 2013; Pardos, Bergner, Seaton & Pritchard, 2013).



Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
YOB	175	1951	1992	1981.84	8.061
Grade	175	.01	1.00	.2668	.35483
EventFrequency	175	31	22118	3120.40	3891.579
DaysActive	175	1	151	27.03	27.988
VideoEvents	175	0	4289	372.88	600.389
Chapter	175	0	18	7.73	5.784
Valid N (listwise)	175				

Table 3. Student Dataset

The mean age of the student population (*n*=175) enrolled in MITs online Electronics and Circuit course for the Fall 2012 session is 30 years of age with a range from 20 to 61 years old (Table 6). A student successfully passing the course has a grade of 55% or above and receives certification denoted in the student table certification field as a “1”. The grade range is between 1-100%. The average final grade for the population is 26%.

Each student had 18 chapters to complete, however, a number of students who completed the course successfully did not finish all chapters which included assignments and exams. The grade for each assignment and exam per chapter was cumulative explaining why a number of students passed without completing 18 chapters. Event frequency and video events is based upon number of clicks within the course module or while viewing video presentations. Video presentations for this course were considered an extra-curriculum activity and was not required in order to pass. They were offered as a supplement to the chapters assigned. Video event frequency average was approximately 373 clicks. Event frequency mean was 3,120 clicks (Table 3).

Only 11% of the class in the sample population were females while 89% accounted for the male population (Table 4). The total number of females who registered for the course was significantly low based upon percentages reported in studies conducted by the National Science Foundation (2012) and the Higher Education Research Institute (2013) where female student enrollment averaged 26% in disciplines related to science, technology, engineering and mathematic (STEM) courses.

Gender

	Frequency	Percent	Valid Percent	Cumulative Percent
female	19	10.9	10.9	10.9
male	156	89.1	89.1	100.0
Total	175	100.0	100.0	

Table 4. Percentages by Gender (Student Dataset)

Of the 19 females enrolled in the course, 26% successfully completed the course. This is significantly higher compared to 22% of the male student population who passed the course (Table 5).

Gender / Certified Cross-tabulation

			Certified		Total
			Did Not Pass	Passed	
Gender	female	Count	14_a	5_a	19
		% within Gender	73.7%	26.3%	100.0%
	male	Count	122_a	34_a	156
		% within Gender	78.2%	21.8%	100.0%
Total		Count	136	39	175
		% within Gender	77.7%	22.3%	100.0%

Table 5. Percentages by Gender Certified / Non-certified (Student Dataset)

A paired sample T-test for event frequency, active days, video events and chapters revealed the average for these four factors was significantly lower for students who did not complete the course in comparison to students who passed the course with a grade of 55% or above (Table 6).

Group Statistics

	Certified	N	Mean	Std. Deviation	Std. Error Mean
EventFrequency	Did Not Pass	136	1587.81	1568.559	134.503
	Passed	39	8464.82	4783.691	766.004
DaysActive	Did Not Pass	136	15.58	13.472	1.155
	Passed	39	66.97	28.902	4.628
VideoEvents	Did Not Pass	136	192.09	243.193	20.854
	Passed	39	1003.33	956.835	153.216
Chapter	Did Not Pass	136	5.13	3.461	.297
	Passed	39	16.82	1.233	.197

Table 6. Averages of Activity in Course Module (Student Dataset)

Seventy-one percent of students between the age of 51 and 61 and approximately 31% between the ages of 40 and 50 years old successfully completed the course with certification (Table 6). This is significantly higher than students between the age of 30 to 39 (15%) and younger students (20%) demonstrated in Table 7. In a similar study investigating MOOC populations, the authors found grades were approximately 12% higher for students over 40 years of age (Guo & Reineke, 2014).

Age Recoded * Certified Cross-tabulation

			Certified		
			Not Certified	Certified	Total
Age Recoded	1951-1961	Count	2	5	7
		% within AgeRecoded	28.6%	71.4%	100.0%
		% within Certified	1.5%	12.8%	4.0%
		% of Total	1.1%	2.9%	4.0%
	1962-1971	Count	9	4	13
		% within AgeRecoded	69.2%	30.8%	100.0%
		% within Certified	6.6%	10.3%	7.4%
		% of Total	5.1%	2.3%	7.4%
	1972-1981	Count	33	6	39
		% within AgeRecoded	84.6%	15.4%	100.0%
		% within Certified	24.3%	15.4%	22.3%
		% of Total	18.9%	3.4%	22.3%
	1982-1992	Count	92	24	116
		% within AgeRecoded	79.3%	20.7%	100.0%
		% within Certified	67.6%	61.5%	66.3%
		% of Total	52.6%	13.7%	66.3%
Total		Count	136	39	175
		% within AgeRecoded	77.7%	22.3%	100.0%
		% within Certified	100.0%	100.0%	100.0%
		% of Total	77.7%	22.3%	100.0%

Table 7. Percentages by Age Certified / Non-certified (Student Dataset)

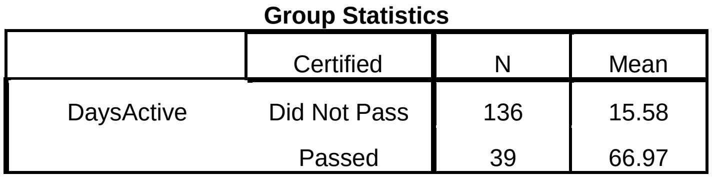
4.1. Model 1

The entire data set (*n*=175) was tested to determine if the predictive modeling system (PMS) was performing as designed. The system functioned as designed and accurately displayed (output) records based upon the embedded SQL formula (*if daysActive <= 27 && yob >= 1982*). If the student's active days for the duration of the course is less than or equal to 27 (average active days for both groups) and the year of birth is greater than or equal to 1982

(30 years of age or younger), records were flagged at-risk “0” in the student table. If the students did not meet this criteria the “at-risk” field was updated with a “1”. Coding aligns with course facilitator’s coding of the certification field in the student’s record.

4.1.1. Active Days

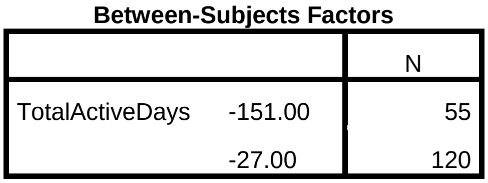
There is a significant difference in the average active days between the two groups who passed the course (67) and did not pass the course (16) as demonstrated in Table 8. As a result, it was decided to use the mean of active days (27) from both groups as a threshold value to predict if a student is at-risk. Results from ANOVA (Table 9) for active days reveals this predictor has an 88% accuracy rate.



Group Statistics

	Certified	N	Mean
DaysActive	Did Not Pass	136	15.58
	Passed	39	66.97

Table 8. Baseline Average (Student Dataset)



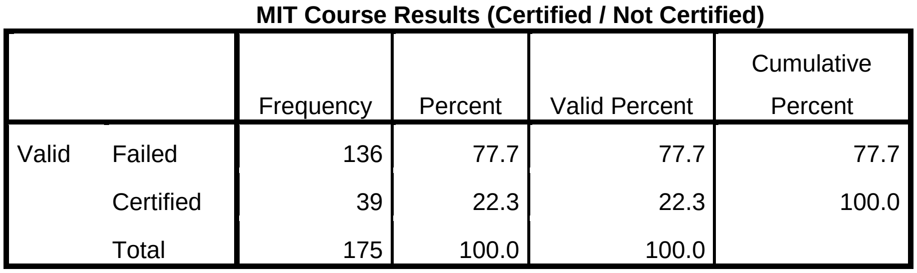
Between-Subjects Factors

		N
TotalActiveDays	-151.00	55
	-27.00	120

Table 9. Active days (Predicted Results)

4.1.2. Year of Birth

In the initial descriptive analysis (Table 7) 67.6% of students between the ages of 20 and 30 years of age was the largest group who did not receive certification. The percentages decrease significantly for students between the ages of 31-40 (24.3%), 41-50 (6.6%) and for students between the ages of 51-61 (1.5%) years of age. Thus, the model predicted with a 67.6% accuracy for this individual factor.

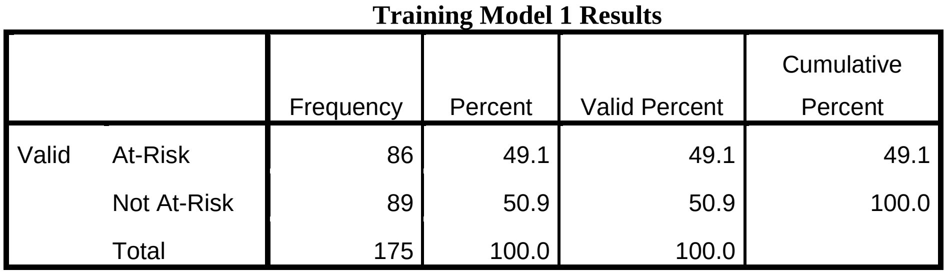


MIT Course Results (Certified / Not Certified)

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Failed	136	77.7	77.7	77.7
	Certified	39	22.3	22.3	100.0
	Total	175	100.0	100.0	

Table 10. Percentages Certified / Failed (Student Dataset)

A total of 78% (136 students) of the 175 who participated in the course did not pass. Only 22% passed the course successfully (Table 10). The PMS model (Table 11) predicted with a 63.2% accuracy rate based upon two combined factors: students who were 30 years of age or younger with active days less than 27 days.



Training Model 1 Results

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	At-Risk	86	49.1	49.1	49.1
	Not At-Risk	89	50.9	50.9	100.0
	Total	175	100.0	100.0	

Table 11. Model 1 Predicted Results

4.2. Model 2

It was determined after the initial model processed to continue validating the entire data set in model two and model three in order to test individual classifiers for accuracy as performed in model one. This inner cross-validation has a dual purpose: model tuning (testing) and identification of the most informative factors in the entire dataset. The holdout procedure / method is then utilized in model four where 60% of student data is reserved for training and 40% is held-out for final testing in model 5.

In the second model the following formula was tested: *($chapters < 11 && $vidView < 373 && $events < 3,120)*. In this formula if chapters completed is less than 11 (and) video view events is less than 373 (and) event frequency is less than 3,120, the students are flagged at-risk. The threshold values are based upon averages obtained during initial analysis of the student dataset (Table 3).

4.2.1. Chapters

The initial threshold value of less than 11 chapters completed was derived from the mean of students who did not pass (5.13 chapters) and students who did pass with an average of 16.8 chapters (Figure 9). However, as Figure 9 demonstrates approximately 98% of the population who did not pass completed up to 14 chapters while 95% of the population that did pass completed 14-18 chapters. A cross-tabulation of chapters when model two was processed revealed a 90.8% accuracy rate for students who are at-risk. Based upon the tabulation results (Table 12), the predictor (chapters) will be adjusted to a value of < = 14 in model four and model five in order to account for the remaining 7.6% who are considered at-risk. The remaining 1% contains two records with missing values.

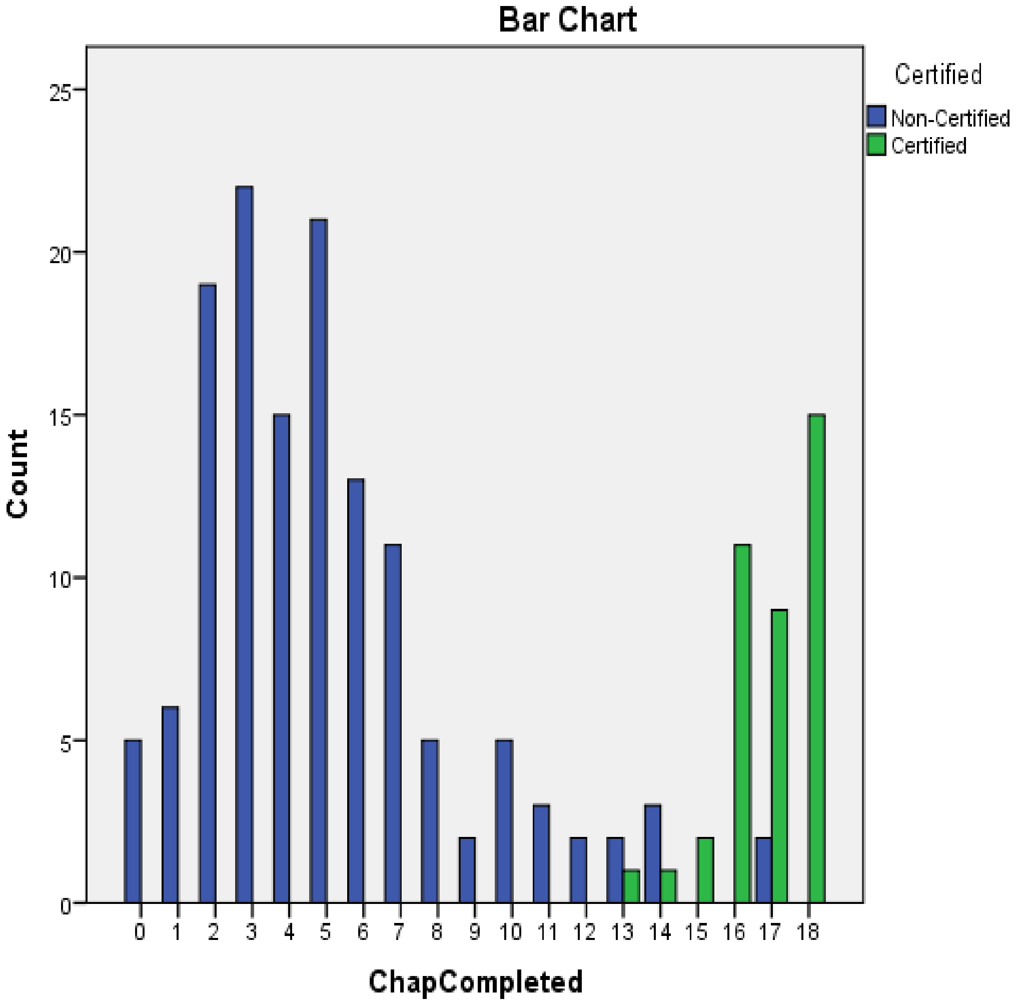


Figure 9. Chapters completed (Student Dataset)

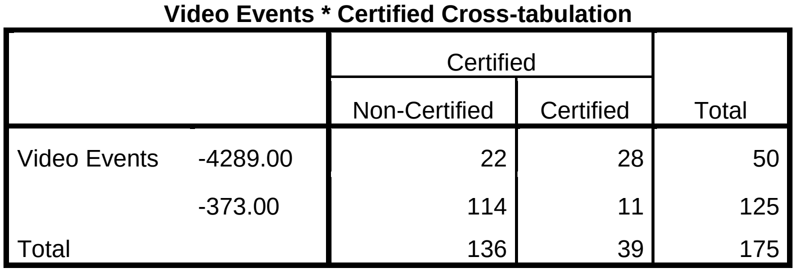
ChapterGroup * Certified Cross-tabulation

			Certified		
			Non-Certified	Certified	Total
ChapterGroup	10.00	Count	119	0	119
		% within ChapterGroup	100.0%	0.0%	100.0%
		% within Certified	90.8%	0.0%	70.0%
		% of Total	70.0%	0.0%	70.0%
	14.00	Count	10	2	12
		% within ChapterGroup	83.3%	16.7%	100.0%
		% within Certified	7.6%	5.1%	7.1%
		% of Total	5.9%	1.2%	7.1%
	18.00	Count	2	37	39
		% within ChapterGroup	5.1%	94.9%	100.0%
		% within Certified	1.5%	94.9%	22.9%
		% of Total	1.2%	21.8%	22.9%
Total		Count	131	39	170
		% within ChapterGroup	77.1%	22.9%	100.0%
		% within Certified	100.0%	100.0%	100.0%
		% of Total	77.1%	22.9%	100.0%

Table 12. Cross-tabulation of chapters (Predicted Results)

4.2.2. Video View Events (clicks)

The mean video events (clicks) for both groups who passed and did not pass the course is 373 (Table 6). Seventy-one percent (125) of the entire student population clicked on video event portion of the course module less than 373 times. Twenty-nine percent (50) of the total population clicked on videos between 373 and 4,289 times (Table 13).

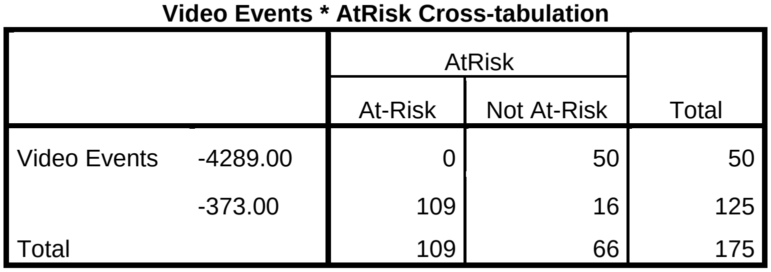


Video Events * Certified Cross-tabulation

		Certified		Total
		Non-Certified	Certified	
Video Events	-4289.00	22	28	50
	-373.00	114	11	125
Total		136	39	175

Table 13. Cross-tabulation Video Events (Student Dataset)

Eighty percent of the students who did not pass the course was identified as at-risk (Table 14). However, the conjunctive formula excluded a number of students who did not fit the remaining criteria for students who participated in the course. Students had more than 3,120 event clicks or completed more than 11 chapters for the course.



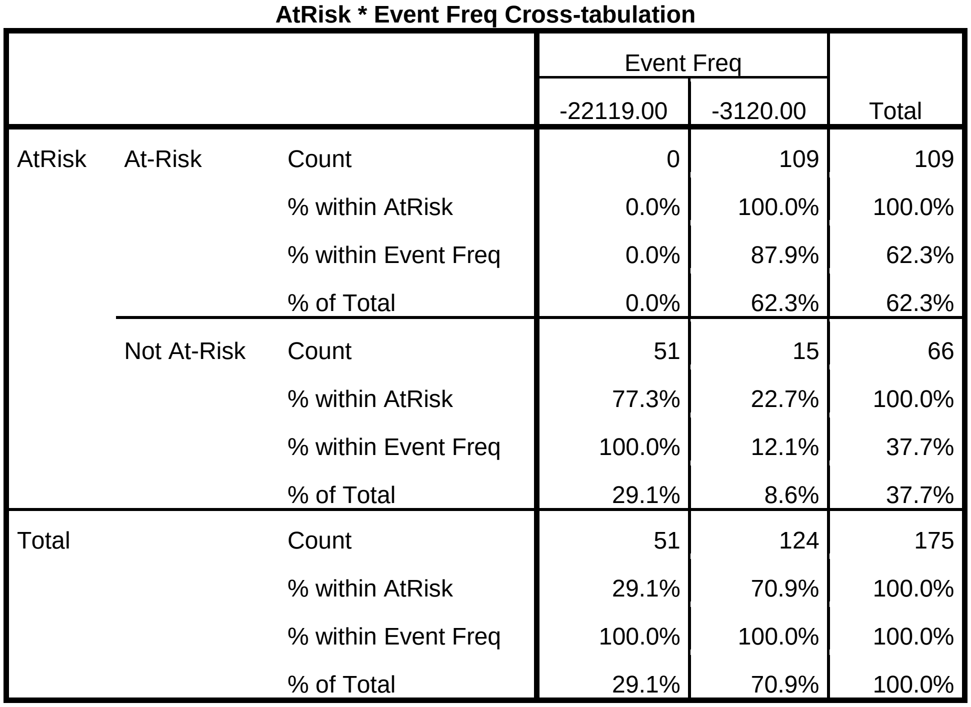
Video Events * AtRisk Cross-tabulation

		AtRisk		Total
		At-Risk	Not At-Risk	
Video Events	-4289.00	0	50	50
	-373.00	109	16	125
Total		109	66	175

Table 14. Video Events (Predicted Results)

4.2.3. Event Frequency

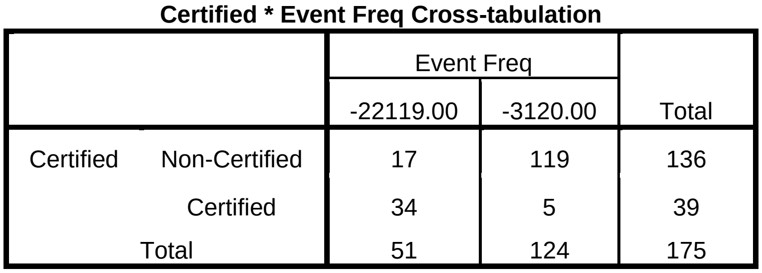
The average event clicks between students who passed and failed is 3,120 with a range of 31 clicks to 22,118 clicks (Table 3). Approximately 88% of the students who did not pass was predicted at-risk (Table 15). Prediction of event frequency is 100% accurate when compared to analysis results of the student dataset (Table 16).



AtRisk * Event Freq Cross-tabulation

			Event Freq		
			-22119.00	-3120.00	Total
AtRisk	At-Risk	Count	0	109	109
		% within AtRisk	0.0%	100.0%	100.0%
		% within Event Freq	0.0%	87.9%	62.3%
		% of Total	0.0%	62.3%	62.3%
	Not At-Risk	Count	51	15	66
		% within AtRisk	77.3%	22.7%	100.0%
		% within Event Freq	100.0%	12.1%	37.7%
		% of Total	29.1%	8.6%	37.7%
Total		Count	51	124	175
		% within AtRisk	29.1%	70.9%	100.0%
		% within Event Freq	100.0%	100.0%	100.0%
		% of Total	29.1%	70.9%	100.0%

Table 15. Cross-tabulation Event Frequency (Predicted Results)



Certified * Event Freq Cross-tabulation

		Event Freq		
		-22119.00	-3120.00	Total
Certified	Non-Certified	17	119	136
	Certified	34	5	39
Total		51	124	175

Table 16. Event Frequency (Student Dataset)

Model two had an overall accuracy rate of 80% based upon combined results from three criteria (Table 17) when compared to the student dataset (Table 18).

AtRisk

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	At-Risk	109	62.3	62.3	62.3
	Not At-Risk	66	37.7	37.7	100.0
	Total	175	100.0	100.0	

Table 17. Model 2 Predicted Results

Certified

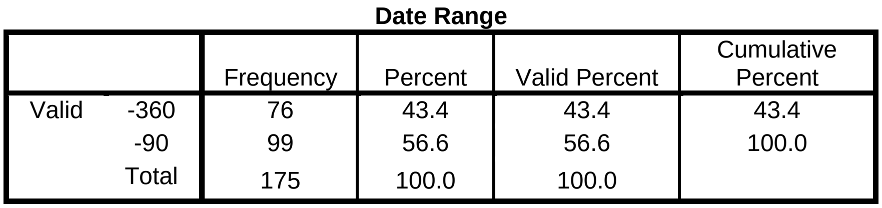
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Non-Certified	136	77.7	77.7	77.7
	Certified	39	22.3	22.3	100.0
	Total	175	100.0	100.0	

Table 18. Student Dataset

4.3. Model 3

The third model examined if a pattern existed between a student's start date and the student's last activity date using a threshold value of 90 days. In this formula, a student is considered at risk if they are not actively engaged in the course for less than 90 days. Fifty-seven percent (99 students) of the entire population participated in the course for less than three months while the remaining 43% were actively engaged in the course between three to twelve months (Table 19). Course start date was September 5, 2012 and course end date December 25, 2012. However, through analysis the data revealed the course was open for the duration of a year based upon the population's last activity date ending September 1, 2013. The data also revealed when a student registered for the course which opened July 24, 2012, they could actively participate in the course beginning on the first day of registration.

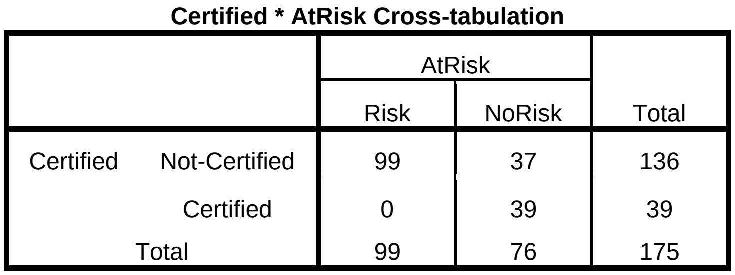
To determine the total months the student continued in the course, the formula entailed subtracting registration date from last activity date to establish duration of months a student continued in the course (Table 19). Seventy-three percent of the total population was identified at-risk (Table 20).



Date Range

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	-360	76	43.4	43.4	43.4
	-90	99	56.6	56.6	100.0
	Total	175	100.0	100.0	

Table 19. Months in Course (Student dataset)



Certified * AtRisk Cross-tabulation

		AtRisk		
		Risk	NoRisk	Total
Certified	Not-Certified	99	37	136
	Certified	0	39	39
Total		99	76	175

Table 20. Cross-Tab of Student Dataset and Predicted Risk

It was determined to exclude delayed / late enrollment in the course as a risk factor. It was initially assumed that a high percentage of students who did not pass the course registered late or started engaging after the course start date. Contrary to this assumption, 72% of the students who did not pass registered early and began engaging in the course prior to the start date. This factor was eliminated for further testing and training.

4.4. Model 4 Training

The dataset was randomly split into two subsets. Training consisted of 105 student records (60%). Seventy records (40%) were set aside for testing. This strategy is relevant when dealing with a small dataset (*n*=175). Each data point (factor) was analyzed and validated during summary analysis and initial model training (model 1-3) to identify which combination of factors were relevant. Training involved four phases. A summary of analysis for each phase is demonstrated in Figure 10.

4.4.1. Phase 1 Training

In phase one, the following code was embedded in the PMS system: *If age is greater than or equal to 1982 (and) gender = male (and) total days active in the course is less than or equal to 27*. If the three criteria is satisfied, the student is flagged at-risk.

Paired Samples Test

		Paired Differences							
					95% Confidence Interval of the Difference				
		Mean	Std. Deviation	Std. Error Mean	Lower	Upper	t	df	Sig. (2-tailed)
Pair 1	Certified - AtRisk	-.371	.524	.051	-.473	-.270	-7.269	104	.000

Table 21. Phase 1 Training Results

A paired sample *t*-test indicated a significant correlation between the certified and at-risk fields for each student record p < 0.05 (Table 21). The PMS system flagged 41 students at-risk (Table 23) of the 80 students who did not pass the course in the training dataset (Table 22). Based upon the formula, the PMS predicted 51% of the students who did not pass were at-risk.

Student Table

		Frequency	Percent	Valid Percent	Cumulative Percent
Failed Pass	0	80	76.2	76.2	76.2
	1	25	23.8	23.8	100.0
	Total	105	100.0	100.0	

Table 22. Student Data

Predicted Results – Training Phase 1

		Frequency	Percent	Valid Percent	Cumulative Percent
At-Risk	0	41	39.0	39.0	39.0
NoRisk	1	64	61.0	61.0	100.0
	Total	105	100.0	100.0	

Table 23. Predicted Results

4.4.2. Phase 2 Training

The following formula was encoded in the PMS system: *If gender equals male (and) chapters completed are less than or equal to 14 (and) event frequency is less than 3,120.* If these conditions are satisfied the records were flagged at-risk (0).

The paired sample *t*-test revealed a significant correlation ($p < 0.05$) when comparing the certified field with the updated at-risk field (Table 24). Prediction accuracy improved 25% in comparison to phase 1 training as demonstrated in Table 25. Seventy-six percent (61 of 80) of the students in the training dataset was accurately identified as at-risk.

Paired Samples Test

		Paired Differences					t	df	Sig. (2-tailed)
					95% Confidence Interval of the Difference				
		Mean	Std. Deviation	Std. Error Mean	Lower	Upper			
Pair 1	Certified - AtRisk	-.181	.411	.040	-.260	-.101	-4.512	104	.000

Table 24. Phase 2 Training Results

Predicted Results – Training Phase 2

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	0	61	58.1	58.1	58.1
	1	44	41.9	41.9	100.0
	Total	105	100.0	100.0	

Table 25. Predicted Results

4.4.3. Phase 3 Training

In phase three, a combination of factors from the first formula (days active; gender) and the second formula (chapters completed) were coded in the PMS: *If days active <= 27 (and) chapters completed <= 14 (and) gender = male* update the at-risk field with a "0". The results yielded similar results found in phase two training (Table 26). Seventy-five percent of the students were identified at-risk.

Predicted Results – Training Phase 3

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	0	60	58.0	58.0	58.0
	1	45	42.0	42.0	100.0
	Total	105	100.0	100.0	

Table 26. Predicted Results

4.4.4. Phase 4 Training

The results of phase four training accurately predicted 85% students were at-risk based upon the following formula: *If chapters are less than or equal to 14 (and) event frequency is less than 3,120 (and) video events are less than 373.* Sixty-eight (85%) of the 80 students who did not pass the course successfully were identified at-risk (Table 27).

Predicted Results – Training Phase 4

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	0	68	64.8	64.8	64.8
	1	37	35.2	35.2	100.0
	Total	105	100.0	100.0	

Table 27. Predicted Results

The certified and at-risk fields were significantly and positively correlated at the 0.001 level (*r*=0.034, *p* =0.001) (Table 28-29). There was not a significant difference (*m*=.114) between the certified (*m*=.24, SD=.428) and at-risk (*m*=.35, SD=.480) fields as demonstrated in Tables 34-35. The Pearson's *r* for the correlation between the certified field in the student records and the output risk value is 0.711 (Table 30). Correlation is significant at the 0.01 level.

Paired Samples Statistics

		Mean	N	Std. Deviation	Std. Error Mean
Pair 1	Certified	.24	105	.428	.042
	AtRisk	.35	105	.480	.047

Table 28. Phase 4 paired sample t-test

Paired Samples Test

		Paired Differences					t	df	Sig. (2-tailed)
					95% Confidence Interval of the Difference				
		Mean	Std. Deviation	Std. Error Mean	Lower	Upper			
Pair 1	Certified - AtRisk	-.114	.348	.034	-.182	-.047	-3.361	104	.001

Table 29. Phase 4 Training Results

Pearson Correlations

		Certified	AtRisk
Certified	Pearson Correlation	1	.711**
	Sig. (2-tailed)		.000
	N	105	105
AtRisk	Pearson Correlation	.711**	1
	Sig. (2-tailed)	.000	
	N	105	105

Table 30. Correlation is significant at the 0.01 level (2-tailed).

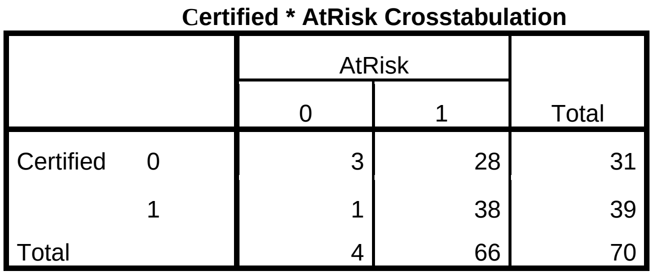
Model Training Summary

Phase	**Formula**	**Student Record Total (Pass)**	**Student Record Total (Fail)**	**Prediction Total (No Risk)**	**Prediction Total (At-Risk)**	**Predictive Accuracy**
1	age>=1982 (+) gender=m (+) actdays <=27	25	80	64	41	51%
2	gender=m (+)chapters<=14(+)events <3,120	25	80	44	61	76%
3	actdays<=27(+)chap<=14(+)gender=m	25	80	45	60	75%
4	chapter<=14(+)events<3,120(+)videv<373	25	80	37	68	85%

Figure 10. Training Results

4.5. Model 4 Testing

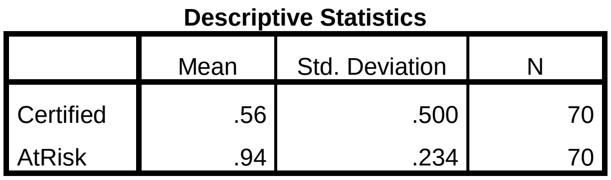
Based upon the results of phase four training (85% predictive accuracy) the formula remained the same for testing: *If days active <= 27 (and) chapters completed <= 14 (and)* *gender = male.* The goal of testing the hold-out set (40%) was to estimate how accurately the predictive model will perform and generalize to the independent dataset. Results of the test did not properly represent the assessment of model performance as expected (Table 31). Cross-tabulation of the hold-out set revealed only 59% students from a population of 70 were accurately identified at-risk or not at-risk when compared to the student data set (Table 32).



Certified * AtRisk Crosstabulation

		AtRisk		Total
		0	1	
Certified	0	3	28	31
	1	1	38	39
Total		4	66	70

Table 31. Cross-tab Results of Test dataset



Descriptive Statistics

	Mean	Std. Deviation	N
Certified	.56	.500	70
AtRisk	.94	.234	70

Table 32. Comparison of mean

The certified and at-risk fields were positively correlated *r*=0.062, *p < .05* (Table 34). There was a significant difference (*m*=.386) between the fields that are certified (*m*=.56, SD=.500) and at-risk (*m*=.94, SD=.234) as demonstrated in Table 38. The Pearson's *r* for the correlation between the certified and at-risk value is 0.152 resulting in a weak correlation (Table 33).

Correlations

		Certified	AtRisk
Pearson Correlation	Certified	1.000	.152
	AtRisk	.152	1.000
Sig. (1-tailed)	Certified	.	.104
	AtRisk	.104	.
N	Certified	70	70
	AtRisk	70	70

Table 33. Pearson correlation

Paired Samples Test

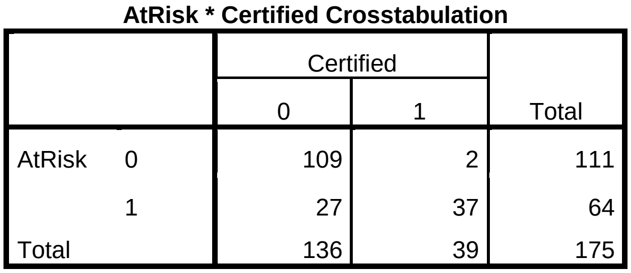
		Paired Differences							
					95% Confidence Interval of the Difference				
		Mean	Std. Deviation	Std. Error Mean	Lower	Upper	t	df	Sig. (2-tailed)
Pair 1	Certified - AtRisk	-.386	.519	.062	-.509	-.262	-6.218	69	.000

Table 34. Phase 4 Testing Results

The cost of the holdout method came in the amount of data that was removed from the model training process. Forty percent (70 records) resulted in significant differences as compared to the final phase of training. As a result, the final model was tested on the entire dataset.

4.6. Final Model

Final model results accurately predicted 80% of the students who were at-risk. One-hundred and nine of the 136 students who did not pass were correctly identified based upon the final formula: *If chapter <= 14 (and) eventfreq < 3,120 (and) videvents < 373* (Table 35).

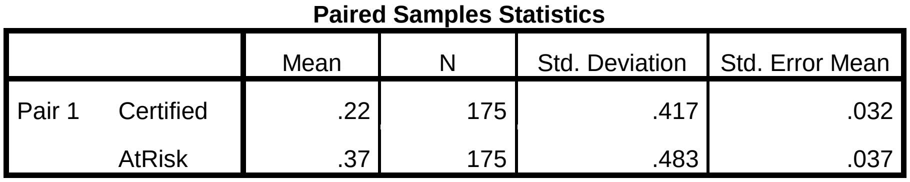


AtRisk * Certified Crosstabulation

		Certified		Total
		0	1	
AtRisk	0	109	2	111
	1	27	37	64
Total		136	39	175

Table 35. Final Predictive Model Results

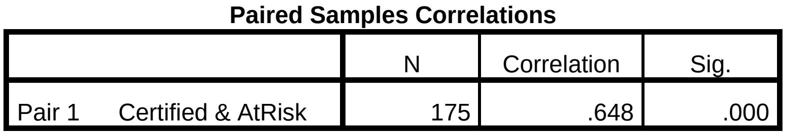
Eighteen percent (32) of the student dataset did not match values in the certified and at-risk field. The mean difference between the two values is .143 (Table 36) and there is a moderate correlation between the two values *r* = .648 with p <.05 (Table 37).



Paired Samples Statistics

		Mean	N	Std. Deviation	Std. Error Mean
Pair 1	Certified	.22	175	.417	.032
	AtRisk	.37	175	.483	.037

Table 36. Final Model Predictive Results



Paired Samples Correlations

		N	Correlation	Sig.
Pair 1	Certified & AtRisk	175	.648	.000

Table 37. Final Model Predictive Results

The relationship between certified and at-risk is positive (.560). Based on the t-value (11.19) and p-value (0.000) there is a positive linear relationship between certified and at-risk fields. A small tolerance value of 1.0 and VIF of 1.0 indicated a linear relationship existed with the independent variables: certified and at-risk (Table 38).

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients			Correlations			Collinearity Statistics	
		B	Std. Error	Beta	t	Sig.	Zero-order	Partial	Part	Tolerance	VIF
1	(Constant)	.018	.030		.596	.552					
	AtRisk	.560	.050	.648	11.197	.000	.648	.648	.648	1.000	1.000

Table 38. Final Model Predictive Results

4.8. Summary of results

The PMS processed a total of twenty-five passes (runs) during model testing and training using various combinations of formulas with all factors in the dataset. The factors included: year of birth, gender, registration date, last activity date, event frequency, days active, span of months active, video event frequency, and chapters completed. When each factor was processed and cross-validated individually: year of birth (age) predicted 68% of the students were at-risk, delayed enrollment (21%), months engaged in the course (73%), gender (males) 90%, days active (87%), event frequency (88%), video events (84%). Registration date, number of months active in the course and delayed enrollment were excluded as predictors, based upon weak predictability results during initial analyses. During training various combinations of factors with strong predictability were combined to test for predictive accuracy. In order to facilitate the validation and interpretation of patterns, group frequencies, paired sample t-tests and regression correlation were used to measure and test the reliability of results produced from PMS output data. The best performance of the model was obtained during the last phase of training (85%). During testing, unexpected results occurred. Using the final formula in training, it was assumed similar rates would occur in testing. The results of testing on the hold-out set (*n*=70) yielded a low accuracy rate (60%). The final formula was then tested across all records (*n*=175) in the dataset resulting in an overall 80% accuracy rate. Results produced by the final model suggests that students in the course who are more actively engaged received certification.

In summary, the resulting analysis of the predictive modeling system addresses key research questions: a.)Which combination of factors when using the PMS system accurately predicts student outcomes? The final formula "*If chapter <= 14 (and) eventfreq < 3,120 (and) videvents < 373"* predicted 80% of the students who were at-risk or not at-risk. b.) Is the PMS easy to use when extracting, analyzing and reporting student outcomes? Selection, updating and extraction of output data was efficient and returned records accurately using a variety of formulas during training and testing. Processing records in the student dataset (*n*=175) and the original dataset in "edX" database system consisting of over 560,000 records returned results instantaneously. Cross-validation of output datasets during all phases resulted in 100% accuracy when imported to analytical software for comparison. The system provides users the option to display and print a list of students who are considered at-risk. The system also provides users the ability to export the entire dataset for further analysis. c.) How easy is it for instructors to modify, maintain and manage the PMS? The system can be easily modified, maintained and managed with a minimal amount of training in order to change factors and formulas for specific programs, courses and student populations.

4.7. PMS Evaluation

The PMS Web-interface was evaluated by three experts in the field of computer science. Evaluation method types were based on a combination of techniques developed by Ivory and Hearst (2001) which focuses on assessment of Web-based systems. Method classes for the evaluation included: testing, inspection, inquiry and analytical assessment (Table 39).

Predictive Model System Evaluation

Method Type	Description
Teaching Method	Online instructions
Co-discovery Learning	Users collaborate
Performance	Measures system performance
Analysis	Analyzes output data
Feature Inspection	Evaluates product features
Usability Inspection	Heuristic evaluation
Standards Inspection	Assess Web compliance standards
Collaboration	Users discuss PMS system
User Feedback	Users submits comments / ranks usefulness
Knowledge Analysis	Evaluate learnability
Design Analysis	Assess design
Programmable	Assess code maintenance of PMS

Table 39. Assessment Methods

A set of test steps (actions), execution conditions and expected results were established for the evaluation in order to determine if the software was working as intended (Appendice A). The evaluators measured performance, analyzed output, assessed the system for Web compliance, and evaluated the design with positive results.

The research questions addressed during the evaluation included: (a) Does the PMS allow a user to easily access the system? During evaluation of the system, each user was provided with a username and password that was non-disruptive while still maintaining privileged restrictions to the application. Permission granting was built into the system in order to minimize unauthorized access to the application and student data. (b) Did the system process in a way that a user expects? The design contained all the features needed to invoke the proper response (output). The system was laid out in a manner that users expected. Users were able to produce accurate results (based upon embedded formulas) via display or print and were satisfied with their ability to export and analyze the entire dataset. (c) Can a user operate the system to a defined level of competence? Users were satisfied with the systems functions and features and reported the system was acceptable from an operational, technological and user standpoint.

Chapter 5

Conclusion

5.1. Discussion

This study reports on the goals and objectives of the development of a predictive modeling system providing a detailed description of the methodology used to design the model in order to predict academic performance in real-time. The motivation for the development of the tool is based upon previous efforts in the research community to design and implement automated systems which identifies students who are at-risk. While a number of efforts have been successful, implementations have proven to be complex and costly to modify and maintain.

The predictive modeling system was designed to process on-demand providing instructors with a tool to monitor student progress in real-time. The PMS can process over relational database management systems (RDBMS) that typically resides on an SQL server. The system subsequently transforms data into an informative risk level report on a selected individual student or an entire class utilizing concepts of inductive logic programming (ILP) using SQL declarative queries. The system provides real-time feedback which relies on multiple factors or and individual factor within an LMS and demographic, academic and financial databases. There exists an implicit relationship between the model and the data that is selected, extracted and reported. PMS formulas and the corresponding risk variables can be easily modified to accommodate a specific program of study, individual course, or characteristics unique to a student population. Another added feature is that individual instructors can use a customized model from past courses on new student data. This eliminates the need to retrain the model.

Another significant element of the PMS is the promptness of the proposed method. The tool provides instructors the ability to identify students before a student decides to drop a course or for a student who may be struggling while enrolled in the course. This enables instructors to facilitate prompt intervention to assist students to continue and successfully pass the course. Early intervention at the course level is a key feature of the model.

The objective for this study is based upon a need to extend on current approaches by introducing an alternative solution to identify students at-risk early in course progression by developing a tool that is portable, easy to use and cost efficient to maintain. The overarching goal is to advance our understanding of low retention rates for non-traditional students enrolled in online programs.

5.2. Implications

This study has important implications for practice within the educational field along with theoretical and research implications relating to the knowledge gained from this research study. This knowledge can be used by researchers and educators to confirm the efficacy of existing predictive mechanisms in place, to modify existing models, and to improve the design of new models.

5.2.1. Theoretical Implications

Theories of retention (Tinto, 1999) and seminal works on retention supported the development of the model. As early as 1993, Tinto acknowledged the importance of institutions improving retention by utilizing analytical methods to examine multiple factors that may contribute to a student dropping out. The combinations of factors which ultimately impacts academic performance include: background characteristics, individual attributes, interactions with peers, faculty and context, and goal commitment.

This study revealed if a student demonstrated a high level of active engagement within the course module (measured by completed chapters, event and video frequency) the frequency of activity was highly correlated to student performance. Engagement has also been positively linked to persistence (Bigatel & Williams, 2015; Lehman & Conceicao, 2011; Rabe-Hemp, Woollen, & Humiston, 2009; Watwood, Nugent, & Deihl, 2009). With the absence of interaction with faculty and staff, students enrolled in the MOOC course under investigation, it was evident from the results that these independent learners success in the course was highly contingent on self-regulation. According to DeBoer, Stump, Seaton, Ho, Pritchard, & Breslow (2013) investigating a similar MOOC population found cognitive, affective and behavioral factors such as interest, self-efficacy, employing effective learning strategies, and satisfaction impact success of a student learning in this environment. Design and messages instructors convey within course content also would have significant implications for student motivation and persistence (Urdan & Schoenfelder, 2006; Wolters, 2004).

The primary intention of this study is to develop a predictive model based upon firmly established theories of retention with a focus on online learning. With MOOCs as a new learning modality a number of theoretical perspectives are emerging to better understand high non-completion rates for students studying in these new online environments. New theories addressing retention of students enrolled in MOOCs is emerging which include chaos theory (deWaard, Abajian, Gallagher, Hogue, Keskin, Koutropoulos, & Rodriguez (2011), connectivist theory (Kop, 2011) and cognitive-behaviorist theory (Rodriquez, 2012). However, diversity in theoretical perspectives often leads to diversity in how courses are designed, developed and delivered.

As in standard online courses that are student-centered, the results of this study suggest that students who are better able to self-regulate, are motivated, and possess a wide range of learning strategies are more actively engaged with course module content. These findings align with previous theoretical views on retention. However, Gašević, Kovanović, Joksimović & Siemens (2014) and DeBoer, et. al, (2013) examining MOOC populations suggest a call for additional studies to explore self-regulatory behaviors arguing that low levels of support and interaction with faculty requires a deeper understanding on a students' ability to self-regulate in this new modality.

5.2.2. Implications in the Research and Educational Field

Given the expansion on development of predictive modeling tools geared towards improving retention for online programs, applying one system to the general population of learners assumes that all students have the same characteristics when identifying risk. Many risk factors used to identify one population may not be applicable to risk factors in another population. For example, in this study, it was initially assumed based upon extant literature and theoretical perspectives of students in STEM courses that females enrolled in the MIT Circuits and Electronics course, under investigation, would be at-risk. However, this study revealed that a higher percentage of female students successfully completed the course. This was significantly higher than male counterparts who passed the course.

PMS processes and subsequent analysis of the original "edX" data source and the Circuits and Electronics course revealed that engagement factors were key predictors of student success. However, these findings for the student population registered in a massive open online course (MOOC) may not apply to other student populations.

Developing a predictive model that can be tailored to specific programs, courses and student populations would be beneficial to all stakeholders. A number of automated at-risk systems at various universities are static generalizing risk predictors across a wide range of student populations. The process of modifying these systems can be viewed as a complex undertaking accomplished by a number of experts. The implications for having a system that can be easily customized to target a specific program, course, student population or one that aligns with an instructors pedagogical methods and practices is an important next step to improve retention.

5.3. Recommendations

The practical recommendations of this study are two-fold: first, current educational database systems (RDBMS) are capable of natively supporting a predictive model that seamlessly integrates Web processing, application programming interfaces (APIs), and query processing for selection, extraction and reporting. Predictive models should be easy to manage and maintain, easy to use and portable across all systems. This study contends that a model and the associated predictive mechanisms containing student factors and threshold values should be customizable for each specific program or course based upon the unique characteristics of the student population. Secondly, with the emergence of research in predictive analytical modeling, there is a need to base the development of said models on established theories and practices that clearly understands the underlying reasons why a student succeeds or does not succeed. By adopting established theory as the basis of a predictive modelling research, there will be less need to create complex algorithms and formulas that require countless iterations in order to produce accurate results to identify students who are at-risk. Finally, current systems in use may be considered pre-defined models. This places limitations on the ability of the system to adapt to changing and diverse student populations. Currently, university wide predictive modeling systems, in place, fits all students to the model rather than fitting the model to specific student populations.

5.4. Summary

In this study, the main goal was to demonstrate the potential benefits for adopting an alternative method to identify students who may be at-risk. The objective was to design a tool that is adaptive and useful to assist instructors in monitoring student progress and identifying if a student is at-risk early in course progression. The findings of the study demonstrates that the PMS can provide timely and automated prediction. While a number of systems excel in predictive performance, this study demonstrates that it is feasible to translate these complex systems to intelligent systems that can be used and managed in everyday practice. The model developed in this study is an adaptive system allowing instructors the opportunity to modify the design, variables and risk formulas easily to match attributes to a specific course or a specific student population. This study also makes a contribution to the understanding of students enrolled in a MOOC where patterns of student engagement emerged as indicators of student success or risk.

References

ACT (2011). National collegiate retention and persistence to degree rates. Retrieved from http://www.act.org/research/policymakers/pdf/retain_2011.pdfT

Agudo-Peregrina, A. F., Hernandez-Garcia, A., & Iglesias-Pradas, S. (2012). Predicting academic performance with learning analytics in virtual learning environments: A comparative study of three interaction classifications. *Computers in Education (SIIE), 2012 International Symposium*, 1-6.

Agudo-Peregrina, A. F., Iglesias-Pradas, S., Conde-Gonzalez, M. A., & Hernandez-Garcia, A. (2014). Can we predict success from log data in VLEs? Classification of interactions for learning analytics and their relation with performance in VLE-supported F2F and online learning. *Computers in Human Behavior, 31*, 542-550.

Aitken, N. D. (1982). College student performance, satisfaction and retention: Specification and estimation of a structural model. *The Journal of Higher Education*, 32-50.

Akçapınar, G., Coşgun, E., & Altun, A. (2013). *Mining Wiki Usage Data for Predicting Final Grades of Students*. Retrieved from http://www.ontolab.hacettepe.edu.tr/wp-content/Publications/Article2013_MAC201310050.pdf

Akdere, M., Cetintemel, U., Riondato, M., Upfal, E., & Zdonik, S. B. (2012). Learning-based query performance modeling and prediction. In Data Engineering (ICDE), *2012 IEEE 28th International Conference*, 390-401.

Allen, I. E., & Seaman, J. (2013). *Changing Course: Ten Years of Tracking Online Education in the United States*. Sloan Consortium, 1-26.

American Institute of Research (2009). Community college dropouts cost taxpayers nearly 4 billion dollars: Low compensation rates generate growing costs to states. Retrieved from http://www.air.org/reports-products/index.cfm?fa=viewContent&content_id=1497

Anaya, A. R., & Boticario, J. G. (2011). Application of machine learning techniques to analyse student interactions and improve the collaboration process. *The Internet and Higher Education, 38*(2), 1171-1181.

Aragon, S. R., & Johnson, E. S. (2008). Factors influencing completion and non-completion of community college online courses. *The American Journal of Distance Education*, *22*(3), 146-158.

Arnold, K. E., & Campbell, J. P. (2013). *U.S. Patent No. 8,412,736*. Washington, DC: U.S. Patent and Trademark Office.

Arnold, K. E., & Pistilli, M. D. (2012). Course signals at Purdue: Using learning analytics to increase student success. In Proceedings of *The 2nd International Conference on Learning Analytics and Knowledge, ACM,* April (2012), 267-270.

Aud, S., & Wilkinson-Flicker, S. (2013). *The Condition of Education 2013*. Government Printing Office. Retrieved from http://nces.ed.gov/pubs2013/2013037.pdf

Ayán, M. N. R., & García, M. T. C. (2008). Prediction of university students' academic achievement by linear and logistic models. *The Spanish journal of psychology*, *11*(01), 275-288.

Baker, R. S., & Yacef, K. (2009). The state of educational data mining in 2009: A review and future visions. *Journal of Educational Data Mining*, *1*(1), 3-17.

Barber, R., & Sharkey, M. (2012). Course correction: Using analytics to predict course success. Proceedings of the *2nd International Conference on Learning Analytics and Knowledge,* (2012), 259-262.

Bean, J. P., & Metzner, B. S. (1985). A conceptual model of nontraditional undergraduate student attrition. *Review of Educational Research*, *55*(4), 485-540.

Bean, J. P., & Metzner, B. S. (1987). The estimation of a conceptual model of nontraditional undergraduate student attrition. *Research in Higher Education*, *27*(1), 15-38.

Berger, J. B., & Lyon, S. C. (2005). Past to present: A historical look at retention. *College student retention: Formula for student success*, *1*, 1-30.

Betts, K., & Lynch, W. (2009). Online education: Meeting educational and workforce needs through flexible and quality degree programs. *I Journal: Insights into Student Services*. Retrieved from http://www.ijournalccc.com/articles/node/72

Bigatel, P., & Williams, V. Measuring Student Engagement in an Online Program. Amazon AWS Systems. Retrieved from http://www.westga.edu/~distance/ojdla/summer182/bigatel_williams182

Black, E. W., Dawson, K., & Priem, J. (2008). Data for free: Using LMS activity logs to measure community in online courses. *The Internet and Higher Education*, *11*(2), 65-70.

Blair, B. S. (2013). Babson research study: More than 6.7 million students learning online. Retrieved from http://www.babson.edu/news-events/babson-news/pages/130107-2012-survey-of-online-learning-results.aspx

Bozick, R., & DeLuca, S. (2005). Better late than never? Delayed enrollment in the high school to college transition. *Social Forces*, *84*(1), 531-554.

Brown, S. M. (2002). Strategies that contribute to nontraditional/adult student development and persistence. *PAACE Journal of Lifelong Learning, 11*, 67-76.

Bunn, J. (2004). Student persistence in a LIS distance education program. *Australian Academic Research Libraries, 35*(3), 253-270.

Calcagno, J. C., Bailey, T., Jenkins, D., Kienzl, G., & Leinbach, T. (2008). Community college student success: What institutional characteristics make a difference?. *Economics of Education Review, 27*(6), 632-645.

Caldwell, B., Cooper, M., Reid, L. G., & Vanderheiden, G. (2008). Web Content Accessibility Guidelines (WCAG) 2.0. W3C.

Campbell, J. P., DeBlois, P. B., & Oblinger, D. G. (2007). Academic analytics: A new tool for a new era. *Educause Review, 42*(4), 40.

Carr, S. (2000). As distance education comes of age, the challenge is keeping the students. *Chronicle of Higher Education*. Retrieved from http://chronicle.com/weekly/v46/i23/23a00101.htm

Chen, X., & Carroll, C. D. (2005). First-Generation Students in Postsecondary Education: A Look at Their College Transcripts. Postsecondary Education Descriptive Analysis Report. NCES 2005-171. *National Center for Education Statistics*. Retrieved from http://nces.ed.gov/pubsearch/pubsinfo.asp?pubid=2005171

Coldwell, J., Craig, A., Paterson, T., & Mustard, J. (2008). Online students: Relationships between participation, demographics and academic performance. *Electronic Journal of E-learning*, *6*(1), 19-30.

Connolly, T., & Begg, C. M (2010). Database Systems: A Practical Approach to Design, Implementation, and Management. *International Edition, Fifth Edition, Pearson Education*.

Council of Graduate Schools (2008). *Ph.D, Completion and Attrition: Analysis of Baseline Demographic Data.* Retrieved from https://www.cgsnet.org/phd-completion-and-attrition-analysis-baseline-demographic-data-phd-completion-project

Council of Graduate Schools (2010). *Ph.D. Completion and Attrition: Policies and Practices to Promote Student Success.* Retrieved from http://www.phdcompletion.org/information/executive_summary_student_success_book_iv.pdf

Council of Graduate Schools (2012). *Retention and Completion of Underrepresented STEM Ph.D. Students: Efforts of the University of South Florida Graduate School*. Retrieved from http://www.cgsnet.org/ckfinder/userfiles/files/AM2012_Liller.pdf

Cox, D. R. (1972). Regression models and life tables. *Journal of the Royal Statistical Society 8(34),* 187-220.

de Raedt, L. (1998). Attribute-value learning versus inductive logic programming: The missing links. In *Inductive Logic Programming* (pp. 1-8). Springer Berlin Heidelberg.

de Waard, I., Abajian, S., Gallagher, M. S., Hogue, R., Keskin, N., Koutropoulos, A., & Rodriguez, O. C. (2011). Using mLearning and MOOCs to understand chaos, emergence, and complexity in education. *The International Review of Research in Open and Distributed Learning*, *12*(7), 94-115.

DeBoer, J., Stump, G. S., Seaton, D., Ho, A., Pritchard, D. E., & Breslow, L. (2013, July). Bringing student backgrounds online: MOOC user demographics, site usage, and online learning. In *Educational Data Mining 2013*.

Delen, D. (2010). A comparative analysis of machine learning techniques for student retention management. *Decision Support Systems*, *49*(4), 498-506.

Demetriou, C. & Schmitz-Sciborski, A. (2009) An innovative intervention for students failing to meet academic standards: the Bounce Back Retention Program case study, *Journal of Widening Participation and Life Long Learning*, *11*(3), 28-31.

Department of Education (2014). Department releases new guidelines on protecting student privacy while using online educational services. *U.S. Department of Education.* Retrieved from http://www.ed.gov/news/press-releases/department-releases-new-guidance-protecting-student-privacy-while-using-online-educational-services

Dobbs, R.R., Waid, C.A. & del Carmen, A. (2009). Students' Perceptions of Online Courses: The Effect of Online Course Experience. *Quarterly Review of Distance Education, 10*(1), 9-26.

Dringus, L. P. (2001). Towards active online learning: A dramatic shift in perspective for learners. *The Internet and Higher Education, 2*(4), 189-195.

Dringus, L. & Terrell, S. (2000). An investigation of the effect of learning style on student success in an online learning environment. *Journal of Educational Technology Systems, 28*(3), 231-238.

Dzeroski, S. (2003). Multi-relational data mining: an introduction. *ACM SIGKDD Explorations Newsletter, 5*(1), 1-16.

Dzeroski, S., Cussens, J., & Manandhar, S. (2000). An introduction to inductive logic programming and learning language in logic. *Lecture Notes in Computer Science,* 3-35.

Freckleton, R. P. (2011). Dealing with collinearity in behavioural and ecological data: model averaging and the problems of measurement error. *Behavioral Ecology and Sociobiology*, *65*(1), 91-101.

Fu, T. C. (2011). A review on time series data mining. *Engineering Applications of Artificial Intelligence*, *24*(1), 164-181.

Fürnkranz, J., Gamberger, D., & Lavrač, N. (2012). *Relational Features*. In Springer-Verlag Berlin Heidelberg (Ed.), Foundations of Rule Learning, 95-112.

Fusch, D. (2011). Tackling the retention challenge: Defining and delivering a unique student experience. *Academic Impressions*. Retrieved from http://www.academicimpressions.com/sites/default/files/0411-diagnostic.pdf

Gasevic, D., Kovanovic, V., Joksimovic, S., & Siemens, G. (2014). Where is research on massive open online courses headed? A data analysis of the MOOC Research Initiative. *The International Review Of Research In Open And Distributed Learning*, *15*(5).

Gazza, E. A., & Hunker, D. F. (2014). Facilitating student retention in online graduate nursing education programs: A review of the literature. *Nurse education today*, *34*(7), 1125-1129.

Geisser, Seymour (1993). *Predictive Inference: An Introduction*. Retrieved from http://books.google.com/books?hl=en&lr=&id=wfdlBZ_iwZoC&oi=fnd&pg=PA1&dq=geisser+1993&ots=p-1k8J_2Cn&sig=XFvvd2ASdXEkBMJ6fWhKU1-3SPo#v=onepage&q=geisser%201993&f=false

Graf, S., Ives, C., Rahman, N., & Ferri, A. (2011). AAT: a tool for accessing and analysing students' behaviour data in learning systems. In *Proceedings of the 1st International Conference on Learning Analytics and Knowledge* (pp. 174-179). ACM.

Grubb, W. (1997). The returns to education in the sub-baccalaureate labor market, 1984–1990. *Economics of Education Review*, *16*(3), 231-245.

Guo, H., Viktor, H. L., & Paquet, E. (2013). Reducing the size of databases for multirelational classification: A sub-graph-based approach. *Journal of Intelligent Information Systems*, 1-26.

Guo, P. J., & Reinecke, K. (2014). Demographic differences in how students navigate through MOOCs. In *Proceedings of the first ACM conference on Learning@ scale conference* (pp. 21-30). ACM.

Habley, W. R., & McClanahan, R. (2004). What works in student retention? Four-year public colleges. *ACT, Inc.* Retrieved from http://eric.ed.gov/?id=ED515398

HarvardX-MITx Person-Course Academic Year 2013 De-Identified dataset, version 2.0, created on May 14, 2014. File name: HMXPC13_DI_v1_5-14-14.csv The md5sum for this release (HMXPC13_DI_v2_5-14-14.csv) is: 2b09c674af772d45dae429045cf7acfc. Retrived from https://thedata.harvard.edu/dvn/dv/mxhx

Herrera, O. L. (2006). Investigation of the role of pre-and post-admission variables in undergraduate institutional persistence, using a Markov student flow model. Retrieved from http://www.lib.ncsu.edu/resolver/1840.16/3441

Higher Learning Commission (2012). Guidelines for the evaluation of distance education. Retrieved from http://www.ncahlc.org/Information-for-Institutions/publications.html

Ho, A. D., Seaton, D. T., Reich, J., Nesterko, S. O., Mullaney, T., Waldo, J., & Chuang, I. (2014). 6.00.x Introduction to Computer Science and Programming-Spring 2013 MITx Course Report (MITx Working Paper# 7).

Horn, L. J. (1998). *Undergraduates Who Work. National Postsecondary Student Aid Study, 1996*. US Government Printing Office, Superintendent of Documents. Retrieved from http://eric.ed.gov/?id=ED421042

Horn, L. J., & Carroll, C. D. (1996). *Nontraditional Undergraduates: Trends in Enrollment from 1986 to 1992 and Persistence and Attainment among 1989-90 Beginning Postsecondary Students. Postsecondary Education Descriptive Analysis Reports. Statistical Analysis Report*. US Government Printing Office, Superintendent of Documents. Retrieved from http://eric.ed.gov/?id=ED402857

Hu, Y. H., Lo, C. L., & Shih, S. P. (2014). Developing early warning systems to predict students' online learning performance. *Computers in Human Behavior*, *36*, 469-478.

Huang, S., & Fang, N. (2013). Predicting student academic performance in an engineering dynamics course: A comparison of four types of predictive models. Computers and Education, 61(2013), 133-145.

Hung, J. L., & Zhang, K. (2008). Revealing online learning behaviors and activity patterns and making predictions with data mining techniques in online teaching. *MERLOT Journal of Online Learning and Teaching*, *4*(4), 426-437.

Guruler, H., & Istanbullu, A. (2014). Modeling Student Performance in Higher Education Using Data Mining. In *Educational Data Mining* (pp. 105-124). Springer International Publishing.

Imielinski, T., & Mannila, H. (1996). A database perspective on knowledge discovery. *Communications of the ACM*, *39*(11), 58-64.

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

Ivankova, N. V., & Stick, S. L. (2007). Students' persistence in a distributed doctoral program in educational leadership in higher education: A mixed methods study. *Research in Higher Education*, *48*(1), 93-135.

Jameson, M. M., & Fusco, B. R. (2014). Math anxiety, math self-concept, and math self-efficacy in adult learners compared to traditional undergraduate students. *Adult Education Quarterly*. Retrieved from http://aeq.sagepub.com/content/early/2014/07/08/0741713614541461.abstract

Kadar, R. S. (2001). A counseling liaison model of academic advising. *Journal of College Counseling*, *4*(2), 174-178.

Kantardzic, M. (2011). *Data mining: Concepts, models, methods, and algorithms*. John Wiley & Sons.

Kelly, P., & Strawn, J. (2011). Not just kid stuff anymore: The economic imperative for more adults to complete college. *The Center for Law and Social Policy (CLASP) and The National Center for Higher Education Management Systems (NCHEMS).* Retrieved from http://www.nchems.org/pubs/docs/NotKidStuffAnymoreAdultStudentProfile-1.pdf

Koller, D., Ng, A., Do, C., & Chen, Z. (2013). Retention and intention in massive open online courses: In depth. *Educause Review*, *48*(3), 62-63.

Kop, R. (2011). The challenges to connectivist learning on open online networks: Learning experiences during a massive open online course. *The International Review of Research In Open and Distributed Learning*, *12*(3), 19-38.

Kupczynski, L., Gibson, A. M., Ice, P., Richardson, J., & Challoo, L. (2011). The impact of frequency on achievement in online courses: A Study from a South Texas University. *Journal of Interactive Online Learning*, *10*(3), 141-149.

Kuusisto, F., Dutra, I., Nassif, H., Wu, Y., Klein, M. E., Neuman, H. B. & Burnside, E. S. (2013). Using machine learning to identify benign cases with non-definitive biopsy. In *15th IEEE International Conference on e-Health Networking, Application & Services (HEALTHCOM 2013), Portugal*.

Lauria, E., Baron, J. D., Devireddy, M., Sundararaju, V., & Jayaprakash, S. M. (2012). Mining academic data to improve college student retention: An open source perspective. *LAK '12: Proceedings of the 2nd International Conference on Learning Analytics and Knowledge*, 139-142.

Lee, Y., Choi, J., & Kim, T. (2013). Discriminating factors between completers of and dropouts from online learning courses. *British Journal of Educational Technology*, *44*(2), 328-337.

Lee, P. H., & Shatkay, H. (2006). BNTagger: Improved tagging SNP using Bayesian networks. *Bioinformatics, 22*(14), 211-219.

Leedy, P. D., & Ormrod, J. E. (2010). *Practical research: Planning and design* (9th ed.). Upper Saddle River, NJ: Prentice Hall.

Lehman, R., & Conceicao, S. (2011). Thinking, Feeling, and Creating Presence in the Online Environment: A Learner's Viewpoint. In *World Conference on Educational Media and Technology* (Vol. 2011, No. 1, pp. 3072-3081).

Lima, R., Espinasse, B., Oliveira, H., Pentagrossa, L., & Freitas, F. (2013). Information Extraction from the Web: An Ontology-Based Method Using Inductive Logic Programming. In *Tools with Artificial Intelligence (ICTAI), 2013 IEEE 25th International Conference on* (pp. 741-748). IEEE.

Liu, S. Y., Gomez, J., & Yen, C. J. (2009). Community college online course retention and final grade: Predictability of social presence. *Journal of Interactive Online Learning, 8*(2), 165-182.

Locke, E. A. & Bryan, J. (1968). Goal setting as a determinant of the effects of knowledge of score in performance. *The American Journal of Psychology*, 398-406.

Lopez, M.I., Luna, J. M., Romero, C., & Ventura, S. (2012). Classification via clustering for predicting final marks based on student participation in forums. in *5th International Conference on Educational Data Mining, EDM 2012*. 2012. Chania, Greece.

Luu, T. D., Rusu, A., Walter, V., Linard, B., Poidevin, L., Ripp, R., & Nguyen, H. (2012). KD4v: Comprehensible knowledge discovery system for missense variant. *Nucleic Acids Research, 40*(1), 71-75.

Macfayden, L. P., & Dawson, S. (2010). Mining LMS data to develop an "early warning system" for educators: A proof of concept. *Computers & Education, 54*(2), 588-599.

Marchewka, J. T. (2006). *Information technology project management*. John Wiley & Sons.

Martinez, D. (2001). *Predicting Student Outcomes Using Discriminant Function Analysis*. Retrieved from http://eric.ed.gov/?id=ED462116

McDaniel, C., & Graham, S. W. (1999). *Student Retention in an Historically Black Institution*. Retrieved from http://eric.ed.gov/?id=ED430474

McDonough, P. M. & Fann, A. J. (2007). The study of inequality. In Gumport, P. J. (Ed.), *Sociology of higher education: Contributions and their contexts* (pp. 53-93). Baltimore: The Johns Hopkins University Press.

McNeely, J. H. (1936). Authority of State Executive Agencies over Higher Education. Bulletin, 1936, No. 15. *Office of Education, United States Department of the Interior*.

Milch, B., & Russell, S. J. (2006). First-order probabilistic languages: Into the unknown. In *Proceedings of the International Conference on Inductive Logic Programming,* 10-24, 2006.

Minaei-Bidgoli, B., Kortemeyer, G., & Punch, W. F. (2004). Enhancing Online Learning Performance: An Application of Data Mining Methods. *Immunohematology, 62*(150), 20-30.

Moore, M. G. (1989). Three types of interaction. *American Journal of Distance Education, 3*(2).

Moore, M.G. (1997). Theory of transactional distance. In D. Keegan (Ed.), *Theoretical principles of distance education* (pp. 22-38). London: Routledge.

Murtaugh, P. A., Burns, L. D., & Schuster, J. (1999). Predicting the retention of university students. *Research in Higher Education*, *40*(3), 355-371.

National Center for Education Statistics (2013). The condition of education 2013. *U.S. Department of Education*. Retrieved from http://nces.ed.gov/pubs2013/2013037.pdf

National Center for Education Statistics (2013). Nontraditional Undergraduates. *Institute of Education Sciences*, U.S. Department of Education. Retrieved from http://nces.ed.gov/programs/coe/indicator_cva.asp

National Science Foundation (2006). *Science and Engineering Indicators 2006.* Retrieved from www.nsf.gov/statistics/seind06

National Science Board (2010). *Science and Engineering Indicators 2010.* Arlington, VA: National Science Foundation. http://www.nsf.gov/statistics/seind12/c2/c2s2.htm

National Science Foundation (2014). *Science and Engineering Indicators 2014.* Retrieved from http://www.nsf.gov/statistics/seind14/index.cfm/chapter-2

Nguyen, H., Luu, T., Poch, O. & Thompson, J. D. (2013). Knowledge discovery in variant databases using inductive logic programming. *Bioinformatics and Biology, 7,* 119-131.

Nunley, C. R. (2007). Community colleges may be losing their edge in educating adults. *Chronicle of Higher Education*, *54*(9), 1-4.

Onwuegbuzie, A. J., Witcher, A. E., Collins, K. M., Filer, J. D., Wiedmaier, C. D., & Moore, C. W. (2007). Students' perceptions of characteristics of effective college teachers: A validity study of a teaching evaluation form using a mixed-methods analysis. *American Educational Research Journal*, *44*(1), 113-160.

Pardos, Z., Bergner, Y., Seaton, D., & Pritchard, D. (2013, July). Adapting bayesian knowledge tracing to a massive open online course in edX. In *Educational Data Mining 2013*.

Park, J. H., & Choi, H. J. (2009). Factors influencing adult learners' decision to drop out or persist in online learning. *Educational Technology & Society*, *12*(4), 207-217.

Pascarella, E. T., & Terenzini, P. T. (1980). Predicting freshman persistence and voluntary dropout decisions from a theoretical model. *The Journal of Higher Education*, 60-75.

Pascarella, E. T., & Terenzini, P. T. (2005). *How college affects students*. (Vol. 2). K. A. Feldman (Ed.). San Francisco: Jossey-Bass. Retrieved from https://edocs.uis.edu/Departments/LIS/Course_Pages/LIS301/papers/How_college_effects_students_534-545.pdf

Perry, B., Boman, J., Care, W. D., Edwards, M., & Park, C. (2008). Why Do Students Withdraw from Online Graduate Nursing and Health Studies Education? *Journal of Educators Online*, *5*(1), n1.

Peissig, P. L., Costa, V. S., Caldwell, M. D., Rottscheit, C., Berg, R. L., Mendonca, E. A., & Page, D. (2014). Relational machine learning for electronic health record-driven phenotyping. *Journal of biomedical informatics,* 1-11.

Qiu, Y., Shimada, K., Hiraoka, N., Maeshiro, K., Ching, W. K., Aoki-Kinoshita, K. F., & Furuta, K. (2014). Knowledge discovery for pancreatic cancer using inductive logic programming.

Rabe-Hemp, C., Woollen, S., & Humiston, G. S. (2009). A comparative analysis of student engagement, learning, and satisfaction in lecture hall and online learning settings. *Quarterly Review of Distance Education*, *10*(2), 207-218.

Rodriguez, C. O. (2012). MOOCs and the AI-Stanford Like Courses: Two Successful and Distinct Course Formats for Massive Open Online Courses. *European Journal of Open, Distance and E-Learning*.

Rosenthal, L., London, B., Levy, S. R., & Lobel, M. (2011). The roles of perceived identity compatibility and social support for women in a single-sex STEM program at a co-educational university. *Sex Roles*, *65*(9-10), 725-736.

Rovai, A. (2002). Building sense of community at a distance. *International Review of Research in Open and Distance Learning, 4*(1), 1-9.

Rovai, A. P. (2003). In search of higher persistence rates in distance education online programs. *The Internet and Higher Education*, *6*(1), 1-16.

Royce, W.W. (1970). Managing the development of large software systems: Concepts and techniques. *Proceedings of the WESCON.*

Rubin, J., & Chisnell, D. (2008). *Handbook of usability testing: How to plan, design, and conduct effective tests*. John Wiley & Sons.

Schneider, M., & Yin, L. M. (2011). The hidden costs of community colleges. *American Institutes for Research*, 1-22.

Severiens, S., ten Dam, G. (2012). Leaving college: A gender comparison in male and female-dominated programs. *Research in High Education*, 53, 453–470.

Shapiro, D., Dundar, A., Chen, J., Ziskin, M., Park, E., Torres, V., & Chiang, Y. C. (2012). Completing college: A national view of student attainment rates. Signature [TM] Report 4. *National Student Clearinghouse*. Retrieved from http://nscresearchcenter.org/signaturereport4/

Snyder, T. D., & Dillow, S. A. (2012). *Digest of education statistics 2011*. National Center for Education Statistics.

Spady, W. G. (1970). Lament for the letterman: Effects of peer status and extracurricular activities on goals and achievement. *American Journal of Sociology, 75,* 4.

St John, E. P. (2000). The impact of student aid on recruitment and retention: what the research indicates. *New directions for student services*, *2000*(89), 61-75.

Steyerberg, E. W., Vickers, A. J., Cook, N. R., Gerds, T., Gonen, M., Obuchowski, N. & Kattan, M. W. (2010). Assessing the performance of prediction models: A framework for some traditional and novel measures. *Epidemiology*, *21*(1), 128.

Stratton, L. S., O'Toole, D. M., & Wetzel, J. N. (2007). Are the factors affecting dropout behavior related to initial enrollment intensity for college undergraduates? *Research in Higher Education*, *48*(4), 453-485.

Southern Association of Colleges and Schools (2011). *Guidelines for addressing distance and correspondence educators.* Retrieved from http://www.sacscoc.org/pdf/081705/Guidelines%20for%20Addressing%20Distance%20and%20Correspondence%20Education.pdf

Terrell, S. R. (2002). The effect of learning style on doctoral course completion in a Web-based learning environment. *The Internet and Higher Education*, *5*(4), 345-352.

Terrell, S. R., Snyder, M. M., & Dringus, L. P. (2009). The development, validation, and application of the *Doctoral Student Connectedness Scale*. *The Internet and Higher Education*, *12*(2), 112-116.

Thayer, P. B. (2000). Retaining first generation and low income students. *Opportunity Outlook, 2,* 8.

Tinto, V. (1975). Dropout from higher education: A theoretical synthesis of recent research. *Review of Educational Research, 45,* 89-127.

Tinto, V. (1993). *Leaving college: Rethinking the causes and cures of student attrition.* (2nd. ed.). Chicago: The University of Chicago Press. Retrieved from http://fdc.fullerton.edu/events/archives/2005/05-01/acadforum/Taking%20Success%20Seriously.pdf

Tinto, V. (1999). Taking student retention seriously: Rethinking the first year of college. *NACADA Journal*, *19*(2), 5-9.

Tinto, V. (2006). Research and practice of student retention: what next? *Journal of College Student Retention: Research, Theory and Practice*, *8*(1), 1-19.

Tinto V. (2012). *Completing college: Rethinking institutional action.* Chicago: University of Chicago Press. Retrieved from http://books.google.com/books?hl=en&lr=&id=zMEy9V4BqDAC&oi=fnd&pg=PR5&dq=Tinto+V.+(2012).+Completing+college+rethinking+institutional+action.+Chicago:+University+of+Chicago+Press&ots=K52Z8_h0Ke&sig=9cIH6-hkKWQLVejb0fPbKEzB1bQ#v=onepage&q&f=false

Trasarti, R., Giannotti, F., Nanni, M., Pedreschi, D., & Renso, C. (2012). A query language for mobility data mining. *Developments in Data Extraction, Management, and Analysis*, 23.

U.S. Census Bureau. (2012). Census bureau releases data showing relationship between education and earnings. *U.S. Department of Commerce, 2009*. Retrieved from http://www.census.gov/newsroom/releases/archives/education/cb09-66.html

Urdan, T., & Schoenfelder, E. (2006). Classroom effects on student motivation: Goal structures, social relationships, and competence beliefs. *Journal of School Psychology, 44*(5), 331-349.

Watkins, K. (2013). An improved recommendation model on grade point average prediction and postgraduate identification using data mining. *Advances in Neural Networks, Fuzzy Systems and Artificial Intelligence*, 186-194.

Watwood, B., Nugent, L., & Deihl, W. (2009). Building from content to community: Rethinking the transition to online teaching and learning: A CTE White Paper/B. *Watwood, J. Nugent, William «Bud» Deihl. Virginia Commonwealth University: Center for teaching excellence*.

Wolters, C. A. (2004). Advancing Achievement Goal Theory: Using Goal Structures and Goal Orientations to Predict Students' Motivation, Cognition, and Achievement. *Journal of educational psychology*, *96*(2), 236.

Wood, D., Kurtz-Costes, B., & Copping, K. E. (2011). Gender differences in motivational pathways to college for middle class African American youths. *Developmental Psychology*, *47*(4), 961.

Woodman, R. (2001). *Investigation of factors that influence student retention and success rate on Open University courses in the East Anglia region*. M.Sc. Dissertation, Sheffield Hallam University, UK.

World Wide Web Consortium. (2013). Web Content Accessibility Guidelines (WCAG) 2.4 Retrieved from http://www.w3.org/TR/UNDERSTANDING-WCAG20/navigation-mechanisms.html

Appendix A

Predictive Modeling System Evaluation

Introduction

The purpose of this usability test is to assess the Predictive Modeling System (PMS). The PMS was designed and developed for the purpose of creating an online tool to predict academic performance which is easy to use and maintain. The system can be implemented across multiple platforms and systems and can process on the majority of data types.

Processing time for the dataset for this test which consists of 175 records is instantaneous. The system has been tested on the original (edX) data source of 560,000 records to measure processing time on larger datasets. Processing time on the original data source was approximately a few seconds to select, extract and display records. The system accurately returned records during training and testing using various formulas.

The embedded formula in the system for this evaluation is the final model formula. The formula: "*If chapter <= 14 (and) eventfreq < 3,120 (and) videvents < 373*" accurately predicted 80% of the students who were at-risk or not at-risk. The system is designed to output results via display with an option to print. Users also have the ability to export the entire database for cross-validation and analysis purposes.

The assessment is divided into two sections: testing and evaluation. Evaluation of the system consists of four method classes with associated method types. Evaluation codes equal "1" for unsatisfactory and "2" for satisfactory. The evaluation also includes questions which applies to the research questions within the dissertation. The results of your evaluation, including comments and questions, will be discussed in Chapter 4 results.

I appreciate the time it will take to test the system. Please, contact me if you encounter issues or have questions.

Mary Fonti

PMS Evaluation Instructions

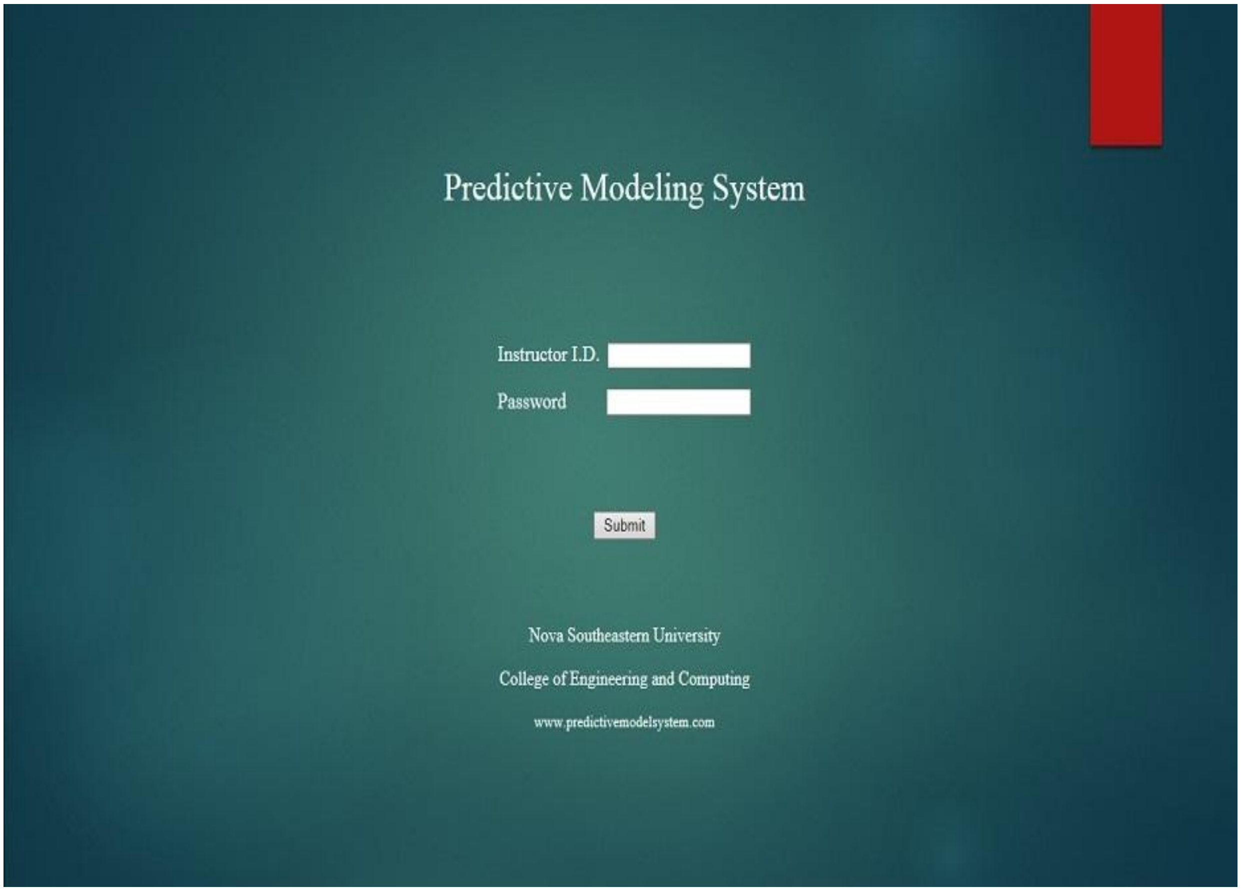
Before the test begins, it is important to clarify the output results that you will see in the display. The Certified field is a static field and part of the student record extracted from the "edX" system. If a "0" is present in the certified field (updated by professors) this indicates that the student failed to complete the course successfully. If a "1" is in the field, this indicates that the student successfully passed the course and received certification. You will then see an At Risk field in the student record. This field was created for the PMS system in order to output a corresponding "0" if the student is considered at-risk or a "1" if the student is not at-risk. This enabled simplified analysis by comparing both fields during the training and testing phases to determine prediction accuracy.

1) *Test*

a.) To log into the PMS system type the following address into your web browser: www.predictivemodelsystem.com. The below screen will be displayed.

b.) Usernames (3): Terrell (or) MacFarland (or) Mukherjee (case sensitive)

Password: NSU-GSCIS (capital letters). The system verifies if the user name or password is correct. If it is invalid, the system will return an error.

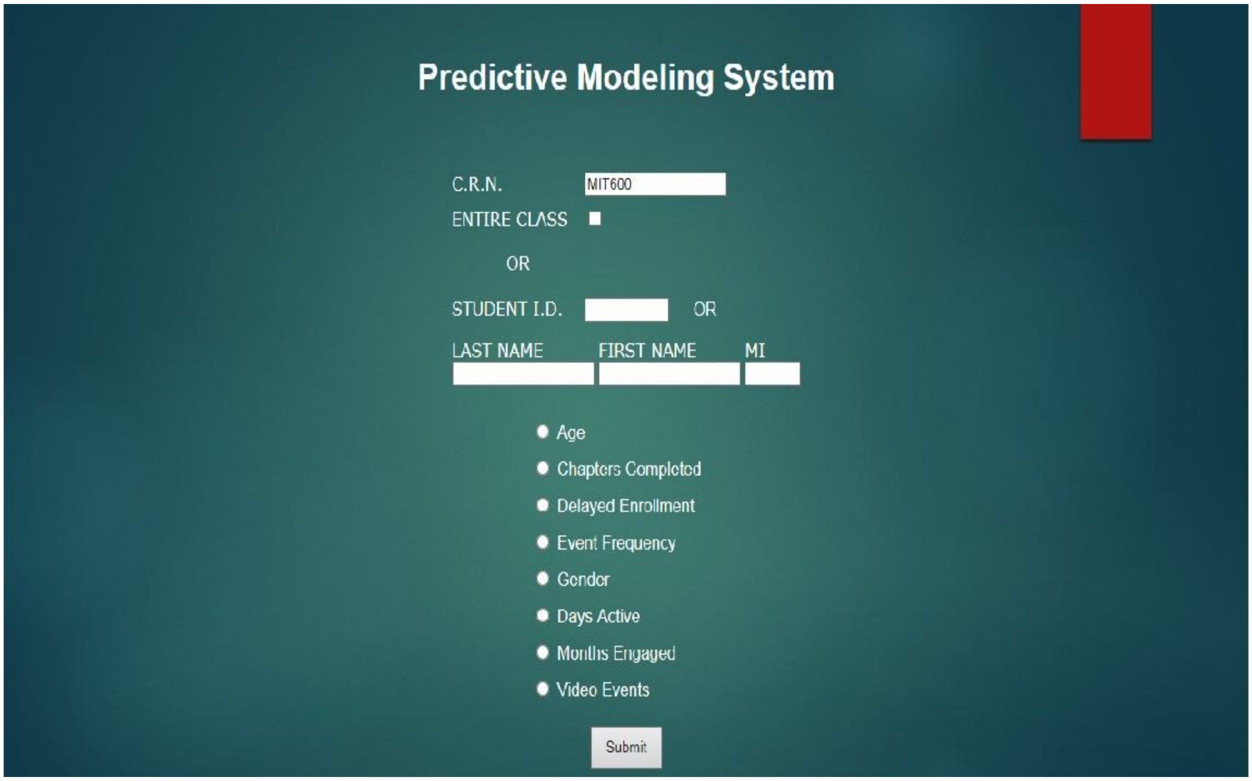

Predictive Modeling System
Instructor I.D.
Password
Submit
Nova Southeastern University
College of Engineering and Computing
www.predictivemodelsystem.com

c.) When you successfully login to the system, the following screen will display.

C.R.N. “MIT600” is the default course registration number.

Click on / check ENTIRE CLASS.

Click the Submit button.

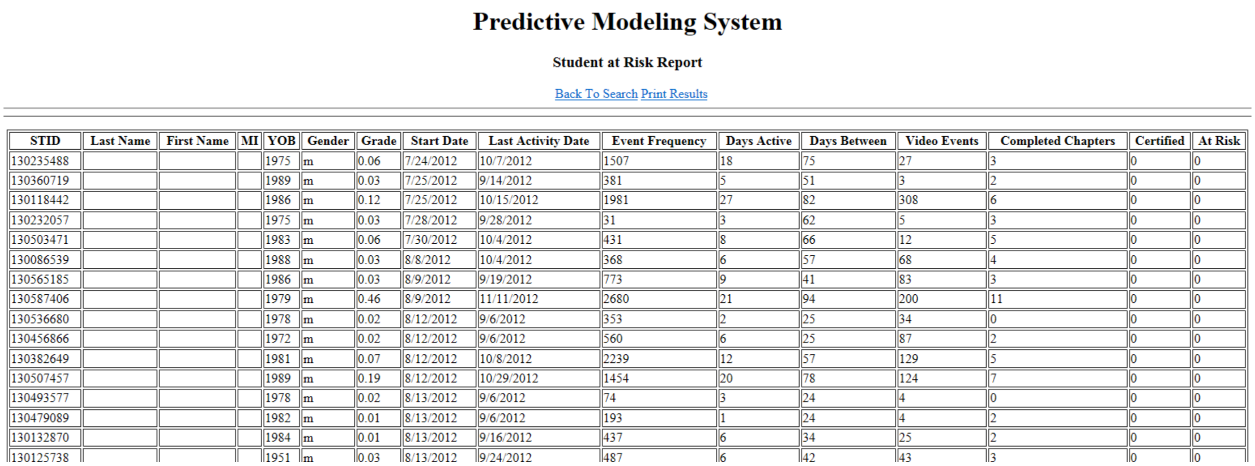


C.R.N. MIT600

d.) When the submit button is clicked the following formula will run:

“chapter <=14 (+) events <3,120 (+) videv <373”

* Note: this is the final models formula. The system will only extract and display student records meeting the three criteria in the formula. A partial output of the display is provided below. This model accurately predicts 80% of the students who are “at-risk.”



Predictive Modeling System

Student at Risk Report

Back To Search Print Results

STID	Last Name	First Name	MI	YOB	Gender	Grade	Start Date	Last Activity Date	Event Frequency	Days Active	Days Between	Video Events	Completed Chapters	Certified	At Risk
130235488				1975	m	0.06	7/24/2012	10/7/2012	1507	18	75	27	3	0	0
130360719				1989	m	0.03	7/25/2012	9/14/2012	381	5	51	3	2	0	0
130118442				1986	m	0.12	7/25/2012	10/15/2012	1981	27	82	308	6	0	0
130232057				1975	m	0.03	7/28/2012	9/28/2012	31	3	62	5	3	0	0
130503471				1983	m	0.06	7/30/2012	10/4/2012	431	8	66	12	5	0	0
130086539				1988	m	0.03	8/8/2012	10/4/2012	368	6	57	68	4	0	0
130565185				1986	m	0.03	8/9/2012	9/19/2012	773	9	41	83	3	0	0
130587406				1979	m	0.46	8/9/2012	11/11/2012	2680	21	94	200	11	0	0
130536680				1978	m	0.02	8/12/2012	9/6/2012	353	2	25	34	0	0	0
130456866				1972	m	0.02	8/12/2012	9/6/2012	560	6	25	87	2	0	0
130382649				1981	m	0.07	8/12/2012	10/8/2012	2239	12	57	129	5	0	0
130507457				1989	m	0.19	8/12/2012	10/29/2012	1454	20	78	124	7	0	0
130493577				1978	m	0.02	8/13/2012	9/6/2012	74	3	24	4	0	0	0
130479089				1982	m	0.01	8/13/2012	9/6/2012	193	1	24	4	2	0	0
130132870				1984	m	0.01	8/13/2012	9/16/2012	437	6	34	25	2	0	0
130125738				1951	m	0.03	8/13/2012	9/24/2012	487	6	42	43	3	0	0

e.) When this screen appears, test all three options found at the top of the page: 1) Print Results 2.) Back to Search 3.) Export student data-set (.csv format) for further examination or cross-validation of results (click the ‘E’ located in upper left hand corner). The entire dataset with results is available for import to analytical software for analysis.

f.) Additional features of the system include the user’s ability to select an individual student. Click on “Back To Search” and enter student id#: 130235488. This will return the student as the record meets formula criteria. However, if you enter student id#: 130589206 results will not be returned as this student is certified and does not meet all three criteria in the formula.

g.) A user may also check individual factors. A formula for each factor has been embedded in the system based upon findings when examining each factor individually. Chapter 2; Section 3; Table 4. I included this function as a demonstration of flexibility and functionality. Please, test individual factors. Return of records will correspond to risk value thresholds found in the following table.

Rank	Field Name	Values	Risk Values
1	Chapters Completed	1 - 18	< 14
2	Event Frequency (Key – Strokes)	31 – 2,218	< 3,120
3	Interaction (Total Days Active)	1-151	< 27
4	Video Events (Clicks)	1-4,289	< 373
5	Gender	M, F	M
6	Delayed Enrollment (Years)	0-38	< 19
7	Age	20 – 61	< = 1982
8	Months Engaged	(1-360)	<=90

2.) Evaluation

Predictive Modeling System Assessment

Method Class	**Method Type**	**Description**	**Comments**	**Evaluation 1: 2**
Testing	Teaching Method	Instructions		
	Learning	Users collaborate		
	Performance	Measures output data		
	Analysis	Analyzes output data		
Inspection	Feature Inspection	Evaluates features		
	Usability Inspection	Heuristic evaluation		
	Standards Inspection	Web compliance		
Inquiry	Collaboration	Users ask questions about PMS		
Analytical	Knowledge Analysis	Evaluate learnability		
	Design Analysis	Assess design		
	Programmable	Assess code		

Please fill out the above form and answer the questions below. A “1” denotes unsatisfactory and a “2” is satisfactory. To access and evaluate PMS code click View on the menu bar then click on Source when the PMS log-in screen or the PMS selection screen is called.

1.) Can a user operate the PMS to a defined level of competence?

2.) Does the PMS process in a way that a user expects?

3.) Is the PMS a valid and reliable tool for predicting student outcomes and monitoring student progress?