USING ADMISSIONS DATA TO CREATE A FIRST-SEMESTER ACADEMIC SUCCESS MODEL

by

David James Byrnes Jr.

A dissertation

submitted in partial fulfillment

of the requirements for the degree of

Doctor of Education in Educational Technology

Boise State University

May 2020

BOISE STATE UNIVERSITY GRADUATE COLLEGE

DEFENSE COMMITTEE AND FINAL READING APPROVALS

of the dissertation submitted by

David James Byrnes Jr.

Dissertation Title: Using Admissions Data to Create a First-Semester Academic Success Model

Date of Final Oral Examination: 12 March 2020

The following individuals read and discussed the dissertation submitted by student David James Byrnes Jr. and they evaluated their presentation and response to questions during the final oral examination. They found that the student passed the final oral examination.

Jesús Trespalacios, Ph.D.	Chair, Supervisory Committee
Jui-long Hung, Ed.D.	Member, Supervisory Committee
Lida Uribe-Flórez, Ph.D.	Member, Supervisory Committee

The final reading approval of the dissertation was granted by Jesús Trespalacios, Ph.D., Chair of the Supervisory Committee. The dissertation was approved by the Graduate College.

ACKNOWLEDGMENTS

I would like to thank Drs. Trespalacios, Uribe-Flórez, and Hung for their continued support throughout my time at Boise State. Each of you has helped me grow not only during this research process but through other endeavors at the college. I truly appreciate the personal touch you have brought to my education even at a distance.

ABSTRACT

Higher education is attracting more students from diverse backgrounds especially at public community colleges. These institutions can help these students attain a quality education at a reasonable price. Unfortunately, community colleges have lower graduation rates than 4-year institutions in part due to the diverse needs and variety in academic preparedness amongst their populations. It can be difficult to identify students most at risk of performing poorly until it is too late. There are multiple ways to predict students' performance. In this study, three common data mining techniques are compared for their accuracy in predicting academic success using only data collected at the point of admissions. Accurate early prediction can allow academic support professionals to intervene and provide intrusive assistance. A neural network model was found to be more accurate than logistic regression and decision tree models. Moreover, data elements of high school GPA, age, and sex were the most important factors in the neural network model.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

CART	Classification and Regression Trees
EDM	Educational Data Mining
EWS	Early Warning Systems
GPA	Grade Point Average
LA	Learning Analytics
LMS	Learning Management Systems
NACADA	National Academic Advising Association
NCES	National Center for Education Statistics
SUNY	State University of New York
WCC	Westchester Community College

CHAPTER ONE: INTRODUCTION

Background of the Study

Higher education in the United States is diversifying in many ways. In 1976, 15.7% of all post-secondary students identified as having a racial or ethnic minority background, but in 2015, that percentage rose to 42.4% (National Center for Education Statistics [NCES], 2016). Additionally, the number of non-traditional students, defined as students who enroll in college after the age of 24 (Hittepole, 2017) is also increasing (NCES, 2014b). Moreover, the number of students accessing higher education online has been steadily increasing (NCES, 2012, 2014a). The students from the three categories mentioned above, bring with them a diverse set of worldviews influenced by unique experiences, academic preparation, and social environments.

Although the landscape of higher education has changed dramatically, it has not changed evenly. Minority, non-traditional, and online students are more likely to enroll in 2-year public community colleges than 4-year counterparts. According to a report from the National Student Clearinghouse Research Center (2017c), more than 155,000 students over the age of 24 who enrolled in college for the first time did so at a 2-year public college, more than the combined new enrollments at all other types of institutions. In terms of ethnicity, 50.8%, 48.5%, and 37.8% of Hispanic, Black, and Asians students, respectively, enrolled for the first time at a community college versus 35.6% of white students who were also more likely than other ethnicities to enroll at a 4-year public institution (National Student Clearinghouse Research Center, 2017b). Meanwhile, a

higher percentage of the students who enroll exclusively in online courses at a public university or college do so at a 2-year school (NCES, 2018). Public education can provide accessible and affordable education, and since more students from diverse backgrounds are choosing community colleges to begin their higher education, it is important for these institutions to help students reach their goals in a timely manner. Since on-time graduation rates for students at community colleges is lower than those for 4-year institutions, it is important for higher education professionals to consider ways to improve successful academic outcomes for their students (National Student Clearinghouse Research Center, 2017a). Successful academic outcomes include students passing their classes with at least minimum requirements to complete their degree and have the ability to transfer credits should they wish to pursue higher degrees.

Statement of the Problem

One way to ensure on-time graduation is by encouraging students to be full-time students, meaning a student is taking 12 or more academic credits per semester (Higher Education Services Corporation, n.d.). Even though taking more classes each semester can expedite graduation, it can be difficult to determine which new students at a community college may struggle with this course load. As noted above, student profiles are becoming increasingly diverse and so are their needs. It is important to provide adequate support services to those who need it most. With continuing students, it is possible to review their academic history at the institution to determine which students might be at risk. However, with new students, it is harder to determine their academic potential. Often the only quantitative information for advisors and academic support professionals to consider are standardized test scores and high school GPAs. Aguinis,

Culpepper, and Pierce (2016) found that standardized test scores are not good predictors of academic success and Vulperhorst, Lutz, de Kleijn and van Tartwijk (2018) found that high school GPAs can be hard to compare among students as different schools have different requirements and academic rigor.

No one piece of information can effectively determine which students are most at risk of struggling with a full-time course load. However, it is possible to use data mining and learning analytics techniques to predict new student performance. Educational Data Mining (EDM) and Learning Analytics (LA) are blossoming fields in which researchers use student data to understand and create solutions for emerging issues (see Calvet Liñán & Pérez, 2015; Siemens & Baker, 2010). EDM and LA techniques have been applied to higher education settings to predict academic success in individual classes (see Baradwaj & Pal, 2011; You, 2016). However, these studies rely on data collected from the classroom environment and present more benefit to individual instructors monitoring their students' progress. There is a lack of literature on how EDM and LA might be used to determine a student's overall risk level (Abu Tair & El-Halees, 2012; Márquez-Vera, Romero, & Ventura, 2013; Zimmermann, Brodersen, Heinimann, & Buhmann, 2015) and fewer articles focus on using this information for the purposes of providing holistic support (Saheed, Oladele, Akanni, & Ibrahim, 2018).

Purpose of this Study

The purpose of this study was to leverage data collected from students at the point of admission to the college to help predict academic outcomes, which may allow academic support professionals to target intervention efforts. More specifically through EDM/LA techniques, I attempted to use prior academic history and demographic factors

most related to academic success to create an academic success model. Output from this model could visualize students' academic risk level in such a way that will be easily accessible by academic support professionals. This sort of model could help identify students who are likely to be successful with a full-time credit course load in their first semester; students who might need additional academic support, such as one-on-one counseling and referrals to other offices. The goal of this study was to determine which demographics and prior academic performance measures help to identify students at risk of performing poorly. The data I examined to achieve this goal are age, credit load, degree program, ethnicity/race, grades, high school information, parents' educational history, residency, sex, socioeconomic status, and standardized test scores. These data elements represent the information that is collected from students via the admissions application or from documents and actions required prior to enrolling.

Significance of this Study

This study is significant as it aims to identify at-risk students before instruction begins. With this knowledge, support professionals can start building meaningful relationships early and arm students with the tools they need to succeed, such as recommending tutoring (Leung, 2015), and helping students learn effective study skills (Wibrowski, Matthews, & Kitsantas, 2017). As noted above, researchers have used EDM to monitor students' progress in individual classes, but this would precede any of those efforts. Since community colleges are generally open-enrollment institutions, there can be a wide range of academic preparedness amongst the students. In order to best focus their efforts, it is important for academic support professionals to easily identify which

students might most benefit from intervention efforts and which students are self-sufficient and not in need of additional or potentially mandated support.

Theoretical Framework

Educational data mining and learning analytics are relatively new fields so there is no established framework from which to approach these research methods. Ranjan and Khalil (2008) proposed a conceptual framework for leveraging admissions data to help counselors make decisions using data mining techniques. In their framework, they noted how it is possible to use data mining in conjunction with professionals' expertise to improve students' academic success. Lei, Yang, and Cai (2017) proposed a model for using EDM for decision making in higher education with many of the same concepts. However, Lei et al.'s (2017) framework also integrated student development theories from Alexander Astin. Astin (1999, 2012) proposed two theories on student development that Lei et al. (2017) combined. The first theory is the input-environment-output model, which suggests that students' demographics (inputs) and environment influence students' success measures (outcomes); inputs also influence environments (Astin, 1999). The second theory is the student involvement theory, which posits that the more a student is involved with a college the more they are likely to succeed (Astin, 2012). Lei et al.'s (2017) framework integrated involvement into the environment piece of the model as students' involvement level influences their perception of an academic environment.

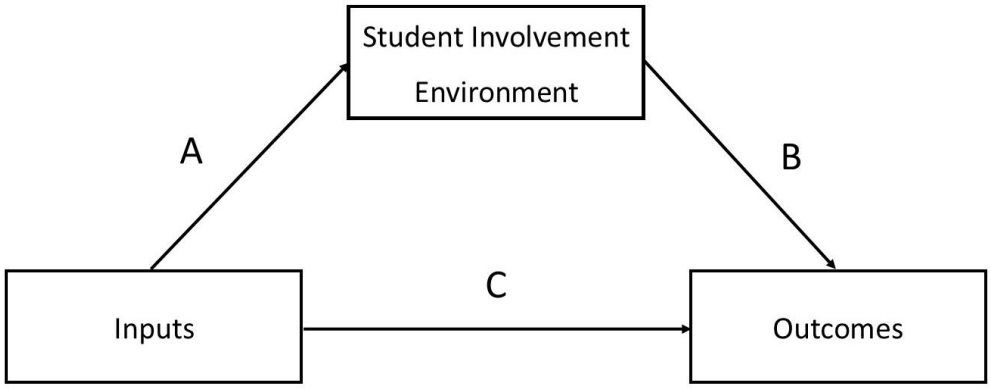

Figure 1.1 Combined student development theory for EDM research from Lei et al. (2017)

Lei et al.'s (2017) framework well suits this study, as it was an investigation into what input variables affect students' outcomes in a specific academic environment and how involving students in academic support might help. Furthermore, this framework relates to the decision-making power of EDM, which is applicable to this study as one of the aims was to develop an at-risk student model. In his original paper on student involvement theory, Astin (1999) wrote:

> Because student personnel workers frequently operate on a one-to-one basis with students, they are in a unique position to monitor the involvement of their clients in the academic process and to work with individual clients in an attempt to increase that involvement. (p. 526)

Therefore, providing these individuals with a tool to help improve success can also be uniquely beneficial.

Additionally, Lei et al. (2017) viewed EDM for decision making as a cyclical process like that of design-based research (Kennedy-Clark, 2013). Through data gathering, processing, and mining, new patterns may arise. These patterns could be useful

data that are then further processed and analyzed. In the field of EDM, it is not always clear which techniques will produce the best results, so there is potential for several iterations of analysis (Abu Tair & El-Halees, 2012). In this study, I explored various data mining techniques to determine which of them produce the most useful results for decision-making.

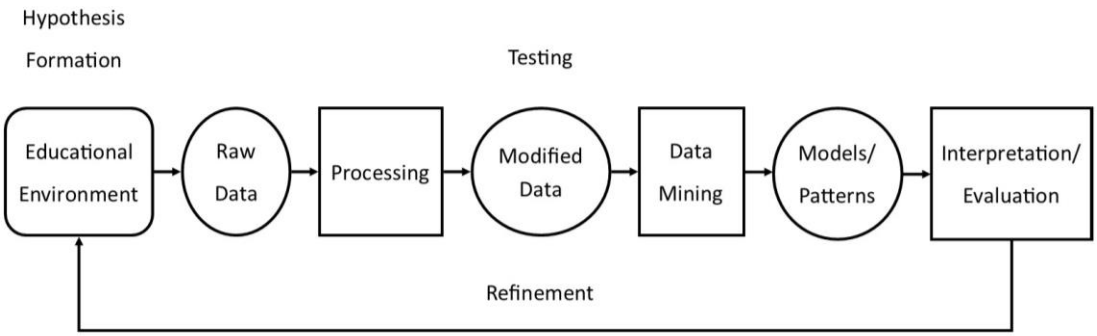

Figure 1.2 Process for extracting, processing and analyzing data to create an EDM model from Lei et al. (2017)

Research Questions

The goal of this study, determining which demographics and prior academic performance measures help to identify students at risk of performing poorly, straddles the line between EDM and LA so it was appropriate to have at least one research question aimed at each area. To answer both research questions, I analyzed variables that are collected from students when they apply to the college, including age, degree program, ethnicity/race, high school information, parents' educational history, residency, sex, socioeconomic status, and standardized test scores. I compared these variables against the recalculated GPA for full-time students, so I also collected information on grades and credit loads.

I ran several prediction models on the variables noted above to answer the first question (Jo, Kim, & Yoon, 2015; Márquez-Vera et al., 2013). The first question concerns the creation of a predictive model, which while applicable to both EDM/LA is more common in LA. I compared how well neural networks, regression analysis, and decision trees predict students' academic success with a full-time course load in their first semester. The method considered most accurate was the one that correctly predicted the most cases (Abu Tair & El-Halees, 2012; Saheed et al., 2018; Singh and Kumar, 2013).

Uncovering patterns is a major focus of EDM, thus the second question about relationships addressed this idea. After I selected the most accurate model, I used feature importance analysis to determine which input variables are most related to the target variable. Therefore, the research questions for this study are:

1. Using data collected at the point of admission, which predictive algorithm generates the best academic success prediction results on the training data set?
2. What key predictors variables are identified by the best predictive model?

Hypotheses

1. $H1_0$: All three data mining/learning analytic techniques used (logistic regression, decision tree, and neural networks) to create a first-semester academic success model based on data collected at the point of admission are equally accurate.
2. $H2_0$: No predictor variables are more important to the most accurate prediction model than other variables

Limitations

This study had the following limitations:

1. The study is limited to one institution, so a model byproduct of this project is not generalizable to other institutions.
2. Students self-reported some of the information on their application, so it was possible that it could be inaccurate or incomplete.
3. Employees entered much of the data into a student information system manually, so there was potential for data entry errors.

Delimitations

As discussed earlier, when determining the risk level for continuing students, it is possible to examine GPA and academic progress. Therefore, this study was limited to analyzing data collected from students at the point of admission for their first semester. Moreover, hundreds of thousands of students have attended the institution that was the site for this study. However, the college has only collected the metrics of interest for this study from students since Fall 2015. Therefore, I only examined records of students entering the college from Fall 2015 until Fall 2018. I chose Fall 2018 as a cutoff, as it is the final semester in which complete data existed prior to the beginning of this study. Finally, I only used information reported to the college via normal processes such as data received from the admissions and financial aid applications. While additional data collected from students may help predict their likelihood of success or level of need, it was outside the scope of this study.

Assumptions

Despite the realistic limitations noted above, some assumptions are required for this study.

1. Data reported by students was accurate.

2. Employees entered data into the student information system correctly unless otherwise demonstrable.
3. The data extracted from the student information system is correct.

While at each of these steps there is a chance for error, it is important to assume the data is correct to complete the analysis.

Definition of Terms

1. Full-time student: a student taking 12 or more academic credits in one semester.
2. Academic success: successfully passing all classes with a minimum GPA of 2.0.
3. Academic support: providing individualized or group support to students that helps them achieve academic success. Support can include one-on-one sessions, referrals to other offices, and tutoring. Professionals who provide this type of support include academic advisors and counselors.
4. Educational data mining: "an emerging interdisciplinary research area that deals with the development of methods to explore data originating in an educational context." (Romero & Ventura, 2010, p. 601).
5. Learning analytics: the use of "sophisticated analytic tools and processes in investigation and visualization of large institutional data sets, in the service of improving learning and education" (Macfadyen & Dawson, 2010, p. 149)

Organization of the Study

I organize the report of this study into five chapters. The first one introduces information on the background, the central problem, purpose, and significance of the study. In addition to those areas, this chapter includes information on the guiding

theoretical framework, research questions and hypotheses, limitations and delimitations and finally assumptions.

Chapter 2 is a literature review of academic counseling and advising, EDM, LA, data visualization, and early warning systems. Chapter 3 focuses on the methodology of the study, which will include a description of the setting, participants, data collection and data analysis.

In Chapter 4, I present the results of the analysis and answer the research questions based on those results. Finally, Chapter 5 includes a discussion of the results and their meaning, as well as recommendations for future study and conclusions.

CHAPTER TWO: LITERATURE REVIEW

Literature from several areas of research informs this study. I will review studies that have used data mining and learning analytics to predict students' academic success and how those studies and specific techniques inform the present research. Since EDM and LA are emerging fields, it is important to review literature in these areas to help guide new research. I will provide an overview of the two fields, then compare, and contrast these fields, highlighting common elements. From there, I will touch on data visualization and early warning systems, and their relation to this study. Finally, I will review articles on the importance of academic counselors and advisors and the potential this research has for improving their work with students.

Related Studies to Predicting Student Outcomes

There is much research on how EDM can be used to predict student performance in individual classes (Baradwaj & Pal, 2011; Yadav, Bharadwaj, & Pal, 2012; You, 2016). From an LA perspective, Pardo, Mirriahi, Martinez-Maldonado, Jovanovic, Dawson, and Gaešvić (2016) created a predictive model for instructors to identify struggling populations of students for additional support. Ranjan and Kahlil (2008) theorized that it is possible for academic counselors to use admission data to help students make decisions about their academic plan through EDM techniques. Since then, many institutions have leveraged this type of data to predict dropout rates for incoming students. In Table 1, I provide details of the data elements, population, and sample size of

studies that have attempted to predict student outcomes other than success in an individual class that inform the current study.

Table 1 Related Studies to Predicting Student Outcomes

Authors	Data Elements Collected	Key Variables	Population	Sample Size
Abu Tair and El-Halees (2012)	Degree program, grades, residency, secondary GPA, sex, type of secondary education	Degree program, residency, sex, type of secondary education	Graduate science and technology students	3,360
Casanova et al. (2018)	Age, credits earned, degree program, GPA, high school GPA, if the university was the student's first choice, parents' educational history, residency, sex	If the university was the student's first choice, sex	First-year undergraduate students	2,970
Delen (2010)	Age, credits earned vs registered, degree program, ethnicity, financial aid need, GPA, marital status, residency, sex, standardized test scores, transfer credits	Credits earned vs registered, financial aid need, GPA	Freshmen undergraduate students	16,066
Marquez-Vera et al. (2013)	Age, family demographics (parents' marital status, number of siblings), previous GPA, scores in specific classes, standardized test scores, student survey data	Family demographics, previous GPA, scores in specific classes, standardized test scores	Secondary students	670
Pal (2012)	Admission type, degree program, caste (social status), economic status, high school GPA, language, parents'	High school GPA, economic status, parents'	Graduate engineering students	1,650

	educational history, parents' occupation, post-secondary GPA, residency, sex	occupation, post-secondary GPA		
Saheed et al. (2018)	Age, degree program, marital status, nationality, parents' occupation, religion, sex, student type, year of entry	Age, degree program, parents' occupation	Undergraduate computer science students	234
Yasmin (2013)	Age, degree program, employment status, marital status, residency, sex, socio-economic status	Employment status, marital status, residency	Graduate distance education students	12,148
Zimmermann et al. (2015)	Age, credits earned vs registered, GPA, scores in specific classes, sex, time to degree completion	GPA, scores in specific classes	Graduate computer science students	171

Casanova, Cervero, Núñez, Almeida, and Bernardo (2018) noted that with the increasing heterogeneousness of college students, demographic factors could help predict the persistence of incoming students. Casanova et al. (2018) found that academic achievement and credits earned are important predictors of students remaining at the institution. Particularly among women, students who had low academic achievement in their first semester were most likely to drop out. Among those who were the highest academic achievers, students who were at a university that was not their first choice were most likely to leave. The authors hypothesize this has to do with those students being able to transfer to another university with their successfully earned credits. This gives

credence to the idea that it is not as important to predict dropout rates among community college students as it is to predict academic success. Students who achieve success may leave the college after transferring, while those who perform poorly in their first semester may not have that option.

Similarly, other researchers have also used a combination of prior academic history and demographic data to predict dropout rates among undergraduate students before they begin (Delen, 2010; Pal, 2012; Yasmin, 2013). Delen (2010) found that some of the most important factors related to student persistence are the ratio of earned credits to registered credits, students' financial aid needs, and their first semester GPA. Predicting the first semester GPA and the successful completion of credits is central to this study. If a prediction model can help determine which students are at risk of performing poorly in their first semester, academic support professionals can intervene and potentially increase student persistence and retention. Additionally, Pal (2012) found that high school GPA and post-secondary GPA were strong predictors of student persistence at the graduate level. This shows how academic achievement at the post-secondary level can have far-reaching consequences for a student's future. They also noted how socio-economic factors such as income and parents' occupation have some connection to student persistence. Yasmin (2013) found that married, employed, remotely located and older students were more likely to drop out than their single, unemployed, urban, and younger counterparts. The groups Yasmin (2013) described meet the criteria of non-traditional students (Hittepole, 2017) who are more likely to enroll at community colleges than 4-year institutions (National Student Clearinghouse Research Center, 2017c). These research studies help to inform the present study, but they all focus on

individual class performance or dropout rates. In a community college, students often stop out for a variety of reasons (Ozaki, 2016) or transfer before graduation (Shapiro et al., 2017). Therefore, predicting whether a student will return may not be as important as predicting if they will successfully complete coursework.

One study conducted at the undergraduate level examined admissions data to predict academic success. Saheed et al. (2018) used student demographics, background information, and academic choices to predict academic success at a university in Nigeria with 95% accuracy. They noted that the student's intended degree program, age, and parents' occupation were effective in predicting academic success. However, their definition of academic success is vague in the paper. To find additional studies that leverage student data to predict overall academic success, it is necessary to look at literature beyond undergraduate institutions. The goal of this study is to determine which demographics and prior academic performance measures help to identify students at risk of performing poorly. That was also the goal of a study by Zimmermann et al. (2015) albeit at the graduate level. These researchers were able to identify several key factors that could predict graduate-level performance. The most important factors found were overall GPA, GPA in the third undergraduate year, and GPA in specific courses. Abu Tair and El-Halees (2012) used similar data and a variety of techniques to find relationships between student admission data and academic performance at the graduate level as well. Through their various data analysis techniques, they found that residency, secondary school type, sex, and degree program all help predict students' final GPA. In a secondary education setting, Márquez-Vera et al. (2013) were able to identify factors that successfully predicted student overall academic success. The static variables most related

to overall success were grades in specific courses, previous year's GPA, standardized test scores, and family demographics. These studies from secondary and graduate-level education will be useful influences on the present study.

Educational Data Mining and Learning Analytics

Educational Data Mining (EDM) is an emerging field in which researchers investigate tools and methods for finding relationships among data to predict outcomes and create models. Many authors have compared this field to Learning Analytics (LA) in which it is possible to use data similarly albeit for slightly different purposes (see Calvet Liñán & Pérez, 2015; Siemens & Baker, 2010). Within both fields, LA and EDM researchers attempt to improve the analysis of large quantities of educational data (Siemens & Baker, 2010). This research orientation represents a data-intensive and data-driven approach to education and its problems (Calvet Liñán & Pérez, 2015; Siemens & Baker, 2010). Additionally, Papamitsiou and Economides (2014) noted that in each discipline researchers gather and process data to reflect on learning processes they are attempting to improve. While conducting reviews of both fields, authors have noted that both EDM (Calvet Liñán & Pérez, 2015) and LA (Siemens, 2012) have the ability to provide feedback in real-time to educators.

The two fields complement one another but have their differences as well. Although these fields are distinct, both disciplines will often use similar approaches, utilize similar methods, and have many of the same goals. I provide an overview of the approaches to using data, data collection and analysis techniques, and aims and goals of EDM and LA in the following paragraphs. A summary of this information is in Table 2.

Approaches to Using Data

These two fields have distinct approaches to how to use data to improve education. In the field of LA, understanding what factors affect and influence educational systems is important (Siemens & Baker, 2010). After gaining a full understanding of these systems, researchers have attempted to leverage human judgment to make improvements (Calvet Liñán & Pérez, 2015; Siemens & Baker, 2010). LA tends to use current methods of analysis or data gathering rather than create new ones (Prakash, Hanumanthappa, & Kavitha, 2014) and may extract this data from social networks (Calvet Liñán & Pérez, 2015) and a wide range of educational activity (Siemens, 2012). Additionally, using data to inform early intervention efforts is common in LA (Siemens, 2012; Prakash et al., 2014). Alternatively, EDM data tends to come directly from software (Calvet Liñán & Pérez, 2015). EDM researchers strive for automatic discovery (Calvet Liñán & Pérez, 2015; Siemens & Baker, 2010) and attempt to find unique data to help solve educational problems (Romero & Ventura, 2013). When working with data, EDM researchers will often need to process the data by normalization and transformation for proper analysis (Calvet Liñán & Pérez, 2015; Romero & Ventura, 2013). Both fields may use predictive modeling to better understand issues in the field of education (Calvet Liñán & Pérez, 2015; Papamitsiou & Economides, 2014; Prakash et al., 2014; Romero & Ventura, 2013; Siemens, 2012; Siemens & Baker, 2010; Sin & Muthu, 2015).

Data Collection and Analysis Techniques

LA and EDM researchers utilize a broad range of techniques for collecting and analyzing data to meet needs and goals. Calvet Liñán and Pérez (2015) noted that LA might employ methods such as concept analysis, sentiment analysis, discourse analysis,

and influence analysis; various authors echoed these claims including Siemens (2012), Siemen and Baker (2012), Romero and Ventura (2013) and Sin and Muthu (2015). The use of these analytical tools fits well with the idea mentioned by Calvet Liñán and Pérez (2015) that LA researchers focused on how to apply data to educational settings. Moreover, a feature of LA is the use of sensemaking models that might provide a better understanding of a learning environment (Calvet Liñán & Pérez, 2015; Romero & Ventura, 2013; Siemens & Baker, 2010). Finally, Siemens (2012) noted that building learner profiles is a common component of LA, but Papamitsiou and Economides (2014) wrote this could be a function of both LA and EDM.

In EDM, researchers may compile data from various repositories and analyze it in a number of ways. Romero and Ventura (2013) noted that EDM programs might retrieve information on student collaboration data, administrative data, demographic data, and data on students' emotions and motivations through methods of processing mining, text mining, and knowledge tracking. They also noted common methods of data analysis are non-negative matrix factorization and outlier detection. Additionally, the methods mentioned by Romero and Ventura (2013) were included in other articles, such as Bayesian modeling (Calvet Liñán & Pérez, 2015; Siemens & Baker, 2010), discovery with models (Prakash et al., 2014; Siemens & Baker, 2010) and classification (Calvet Liñán & Pérez, 2015; Papamitsiou & Economides, 2014). Other methods mentioned were cross-validation (Siemens & Baker, 2010) and distillation (Prakash et al., 2014). Some methods of data analysis were noted by some authors of being primarily tied to EDM while other authors noted their use in both EDM and LA. Siemens and Baker (2010), Romero and Ventura (2013), and Prakash et al. (2014) wrote that clustering and

relationship mining is common in EDM, while Calvet Liñán and Pérez (2015) noted the use of these methods in both fields and Papamitsiou and Economides (2014) stated the use of clustering can be found in both fields. Finally, Siemens and Baker (2010) highlighted the use of student modeling in EDM and again Papamitsiou and Economides (2014) noted the applicability of this method to both EDM and LA.

Other methods are hard to categorize into primarily EDM or LA, so these methods of data gathering and analysis represent the overlap between the two disciplines. For instance, visualization is a method utilized in both fields. Siemens and Baker (2010) and Calvet Liñán and Pérez (2015) both note how visualization is used in EDM, while Sin and Muthu (2015) mentioned its use in LA, but Romero and Ventura (2013) wrote that it may be found in both fields. Papamitsiou and Economides (2014) noted that regression is common in these two disciplines and mentioned that researchers may aim to find the most significant factor when using modeling techniques. Sin and Muthu (2015) echoed their message about the use of both in EDM and LA.

Aims and Goals

EDM and LA share the same overarching goal of using data to improve education (Papamitsiou & Economides, 2014). However similar in approaches and methods, each field has unique goals and some common ones as well. Many authors noted that there was a focus in LA on gathering and analyzing data about learners and their contexts to gain a better understanding of these learners and environments. With this understanding, researchers can attempt to improve the learning environment (Calvet Liñán & Pérez, 2015; Papamitsiou & Economides, 2014; Prakash et al. 2014; Romero & Ventura, 2013; Siemens, 2012; Siemens & Baker, 2010). Through a mix of strategies from computer

science and social science fields such as sociology and psychology (Siemens & Baker, 2010; Sin & Muthu, 2015), LA researchers aim to gain a holistic understanding of educational issues. In LA, Siemens (2012) noted that common goals are to optimize student success, improve advising practices, improve educators' use of technology in learning environments, create personalized and adaptive learning modules, and improve curriculum design. Sin and Muthu (2015) also noted LA's ability to improve curriculum design as well as the application of improving student performance that Calvet Liñán and Pérez (2015) also found. Some other goals of LA mentioned by Calvet Liñán and Pérez (2015) were to improve faculty performance, increase student understanding of topics, improve grading accuracy, assist in instructors identifying teaching strengths, and recommend effective uses of resources. Finally, Siemens and Baker (2010) noted that commonly LA researchers aim to meet the needs of interested stakeholders with the use of data.

However, in EDM, researchers focused on developing methods that may examine unique or new forms of data in hopes of gaining a better understanding of students and their learning settings (Calvet Liñán & Pérez, 2015; Siemens, 2012; Siemens & Baker, 2010; Sin & Muthu, 2015). In this field, researchers will break down data into its components to find granular relationships between learning factors (Calvet Liñán & Pérez, 2015; Papamitsiou & Economides, 2014; Prakash et al., 2014; Siemens, 2012; Siemens & Baker, 2010). In terms of EDM, Calvet Liñán and Pérez (2015) noted the goals of improving machine learning and enhancing scientific research. They also wrote about other common goals which are found in papers from other authors, including automated adaptation (Siemens & Baker, 2010), pattern detection in large data sets

(Romero & Ventura, 2013; Papamitsiou & Economides, 2014), course content reorganization (Romero & Ventura, 2013), model building, and measurement the effect of pedagogical differences (Prakash et al., 2014). Additional goals of EDM mentioned by authors are measuring student performance, determining how to best extract data (Sin and Muthu, 2015), understanding how students learn, creating systems which automatically select content, advancing knowledge about learning theory (Prakash et al., 2014), determining optimal learning resource placement based on usage, creating recommender systems, and understanding how students research and retrieve information individually and as a group (Romero & Ventura, 2013).

Some goals might be primarily associated with one field but also have application in the other, while other goals are very common in both EDM and LA. Siemens (2012) noted that improving student self-awareness is a goal of LA, but Papamitsiou and Economides (2014) stated researches in either field might have this goal. Similarly, Siemens (2012) noted improvement of pedagogical practice is common in LA, but Sin and Muthu (2015) contended that it is used EDM as well as LA. Moreover, Siemens and Baker (2010) and Calvet Liñán and Pérez (2015) noted the common goal in LA research of empowering learners and instructors, while Papamitsiou and Economides (2014) noted the use of this goal in EDM and LA. On the other hand, the use of EDM to identify which learners would benefit from individual feedback and suggestions is common according to Romero and Ventura (2013) and Prakash et al. (2014), but Papamitsiou and Economides (2014) noted the use of this goal in both disciplines. More evenly split between the EDM and LA is the goal of improving assessment, Romero and Ventura (2013) noted its use in EDM, and Sin and Muthu (2015) noted its use in LA, while both Siemens and Baker

(2010) and Papamitsiou and Economides (2014) contended that this goal is common for both fields. Siemens and Baker (2010) wrote that the aim of both EDM and LA is to help improve the understanding of educational problems and improve the selection and planning of intervention efforts. Finally, Calvet Liñán and Pérez (2015) mentioned the importance of improving decision making in both disciplines.

Table 2 Aims, Approaches, Data Collection and Analysis and Goals of EDM/LA

	Educational Data Mining	Learning Analytics	Both EDM/LA
Approaches to using data	Uses data from software. Focuses on automatic discovery. Normalizes or transforms data. Uses unique forms of data.	Uses data to make decisions. Uses current existing methods of analysis. Uses data for early intervention initiatives.	Uses models to predict outcomes.
Data collection and analysis	Uses collation, administration, demographics, motivation and emotional data. Analyses with processing and text mining, and knowledge tracking Uses methods of cross-validation, distillation, modeling, classification.	Uses concept, sentiment, discourse, and influence analysis. Uses sense-making models	Builds learner profiles. Uses clustering, and student and relationship mining. Uses regression and visualization.
Aims and Goals	Examine new forms of data. Understand the underlying relationships between data. Improve machine learning, develop new EDM techniques. Leverage data to understand students, improve educational technology tools and measure performance.	Understand learners and their environments. Understand issues Holistically. Optimize student's academic success, improve technology use, learning, and education	Understand large sets of data. Improve learning processes and provide feedback to educators. Improve self-awareness,

		environments. Meet stakeholders' needs.	pedagogical practices, feedback and assessment, and understanding of problems.

Relevance to the Current Study

Elements from both fields were central to this study as the approaches, methods, and goals for EDM and LA overlap when building a prediction model for use by educators to improve students' academic success. The approach to realizing this aim was using a model to gain a better understanding of how student's backgrounds and prior academic history play a part in their academic success (i.e. Papamitsiou & Economides, 2014). I attempted to create a tool to understand an educational system better and help inform decision-making; this is inherently an LA technique (Siemens & Baker, 2010) and I then examined that tool to identify unique data patterns (Romero & Ventura, 2013). Moreover, the data came from various sources through mining techniques with the hope of creating a sense-making model for academic support professionals (i.e. Romero & Ventura, 2013). In this study, I aimed to use raw student data to understand trends better while also leveraging that data to improve the learning environment for disadvantaged students (i.e. Calvet Liñán & Pérez, 2015). Finally, by achieving the goal of this study, an implication of these results might improve education (Papamitsiou & Economides, 2014) and students' academic success by enhancing advising with useful technology (Siemens, 2012) while also adding to scientific research in the field (Calvet Liñán & Pérez, 2015).

Common EDM/LA Techniques Used

Reviews on LA and EDM show that it is possible to employ a variety of methods and techniques from these fields to discover underlying trends in education (Papamitsiou & Economides, 2014; Peña-Ayala, 2014; Romero & Ventura 2010). Specifically, when used to create models to predict academic performance correctly, researchers have used regression analysis, neural networks, and decision trees. Since any or all of these techniques could accurately help predict students' academic success rates with the data I collected, they all have potential.

Regression

Regression analysis can be either linear or logistic if the variables are continuous or categorical in nature, respectively. Using proxy variables of students' time management skills, Jo et al. (2015) were able to predict students' academic success in an online course. They first ran a correlation analysis to determine the factors best correlated with students' final grades and then ran a linear regression to understand the impact of the different variables. By running a correlation analysis first, they were able to reduce the number of variables in the final analysis. In their final linear regression analysis, they determined that three proxy variables of time management explained 34.7% of the variance in students' academic performance. Rogers, Colvin, and Chiera (2014) compared linear regression to a more simplistic indexing approach for predicting students at risk of failing. The researchers found linear regression to be the better predictor of academic performance, the factors used in the linear regression accounted for 57% of the variance among students' final grade. While a more simplistic method might be preferable to those who are not experts in data analysis, increased accuracy was the focus

of this study. In another study, demographic and prior academic history variables accounted for 54% of the variance in the academic performance of graduate students using a linear regression model (Zimmermann et al., 2015).

Waddington, Nam, Lonn, and Teasley (2016) leveraged logistic regression techniques to improve an early warning system based on resources utilized by students. These authors were able to use student behavior to make accurate predictions of their end of term grades. This allowed academic advisors to intervene with those predicted to do poorly. Students who accessed exam preparation materials via the learning management system (LMS) were more likely to earn an A or B in the class than those accessing lecture materials. In an attempt to classify students as at-risk/not at risk, Macfadyen and Dawson (2010) used logistic regression to predict academic achievement with an accuracy of 73.7%. The authors determined that the model was more likely to incorrectly classify students as at-risk versus not at-risk, which is the preferred error. This type of error was the preferred error in the present study as well since it is better to provide services to students who may not need them than miss students truly at risk.

Decision Trees

Decision trees are classification methods that are visual in nature. Based on a series of if/then criteria, individual cases are categories into a predictive outcome. Professionals can follow the branches of the tree based on a student's record to see where the student risk level. Yadav et al. (2012) compared three decision tree algorithms to create a predictive model of student performance. Of the models they compared, the classification and regression trees model was most accurate with a correct classification rate of 56.25%. In comparing these tools, it was found that some models placed more

students into the correct risk level but when wrong could be very far off. Other models had fewer directly accurate placement of students' risk levels but were closer to the correct level when wrong. This study highlighted yet another technique for prediction worth considering and noted the importance of comparing different methods. In a similar study, Baradwaj and Pal (2011) used decision trees for predicting student performance with success. Although their model had an accuracy rate of 50% for exact prediction, it had an 80% accuracy rate within one letter grade. In both studies, the authors noted how decision trees are also a natural visualization that can assist in early intervention efforts by support staff. This type of byproduct would be useful to the current study, where the final product could be something that practitioners may use to improve their work.

Yasmin (2013) used decision trees in order to predict the dropout rates of online students with an accuracy rate of 84.8%. This study is of note since data used in this study was primarily demographic and focused on a population seen more frequently at the community college level. Similarly, Mohamed and Waguih (2018) used decision trees to build a performance predictor model for academic advisors. Using academic data including high school GPA and degree program, along with demographic information, they were able to predictor performance with over 87% accuracy. Finally, Saheed et al. (2018) used the same type of demographics as Mohamed and Waguih (2018) with an accuracy of over 98%.

Neural Networks

With the number of variables I intend to collect, one option to consider is neural networks since this EDM technique can find hidden connections among large sets of input variables to help predict outcomes. Ramesh, Parkavi, and Ramar (2013) found that

a neural network called multi-layer perception was more accurate than other techniques in predicting student outcomes with the use of demographic data. The researchers correctly predicted 72.4% of the cases with this neural network. Singh and Kumar (2013) noted that the same neural network was among the most accurate techniques they used to predict which factors affected student recruitment; it correctly classified all instances. Furthermore, when comparing techniques to predict student's final grades, Jishan, Rashu, Haque, and Rahman (2015) found neural networks to be the most or among the most accurate approaches depending on the information input. When using demographic data to predict retention, Delen (2010) found neural networks to be nearly 80% accurate. The data collected for Delen's (2010) study is nearly identical to the data I collected.

Data Visualization

Data visualization is a process in which data expert or computer programs depict large amounts of data pictorially or graphically. This information would be nearly impossible to understand and compare in its raw form but with data visualization techniques, individuals may be able to make sense of these large data sets. Since visualizations take large sets of data and represent meaningful connections about information graphically, it has close ties to statistics and cartography (Huff, 1982; Monmonier, 1996). With advances in technology, large repositories of data are more available to those without advanced computing degrees (see Akanmu and Jamaluddin, 2016; Rose 2017). As such, data visualization represents an opportunity for decision-makers in a variety of fields to make use of such data to improve their organizations.

Relevance to the Current Study

In the field of EDM, researchers collect large amounts of data with many variables. This study was no different. I collected data on students' demographics including, age, race/ethnicity, sex, and academic measures such as standardized test scores, degree programs, and high school GPA. Since visual aids can express ideas quicker than written language (Goldsmith, 1984; Hansen, 1999; Tufte, 1983), it was helpful to utilize these tools to express the many group differences that arose from my analysis. Tools that help to create informative and intuitive graphics are useful when infusing visualization into text (Lin, Fortuna, Kulkarni, Stone, & Heer, 2013). Additionally, since one output of the project would be a model for understanding predicting students' academic success, a visualization of that model can help portray complex ideas (Hansen, 1999; Tversky, 2001).

Visual nature of EDM/LA Techniques

EDM takes larges sets of data and analyzes them to find trends. As noted above, these trends are only understandable through some sort of visual representation. Moreover, an aim of LA is to help educators better understand learning environments to make decisions that data visualization has to power to achieve. Some EDM/LA techniques are inherently visual such as decision trees and clustering which create graphic outputs. However, other methods such as regressions and classifications do not have a natural visual component, so it is important to draw from the field of data visualization to improve data models. Classification techniques, as their name suggests, classifies specific cases based on overall data patterns. It is possible to translate the information from classification methods into a visual representation. For instance, student

classifications of high, medium, and low risk, translated into the colors red, yellow, and green is intuitively understandable to anyone familiar with traffic lights. Agnihotri and Ott (2014) understood how important it is for end-users to be able to interpret models easily. When creating their at-risk model for incoming students, they collaborated with counselors so that the output would be easily digestible and useful to their intervention effort.

Using decision trees, Casanova et al. (2018) were able to identify students at risk of dropping out after their first year. As noted above, the results were inherently visual so stakeholders could intuitively interpret these results for decision-making purposes with future students. Xing, Guo, Petakovic, and Goggins (2015) stressed the importance that teachers be able to digest the results obtained from EDM/LA techniques. Using a relatively advanced EDM technique known as genetic programming, these researchers were accurately able to predict student final grades with data on participation. These researches converted the result to a more simplistic if/then rule tree so that instructors could more easily interpret individual cases to provide academic support.

Early Warning Systems

Early warning systems (EWS) aim to identify students who might struggle as early as possible so that educators can intervene. These tools may use information manually entered by instructors, data analyzed through EDM techniques, or both. Creating a model that can predict students' academic success with admission data is a type of EWS, one that can alert support professionals to at-risk students even before classes begin. EWS use databases to decipher patterns and trends among students to

identify which students may be at risk. However, identifying students is not enough; these systems have the potential to help professionals get students back on track.

Studies on EWS

As EWS become more common, researchers have begun to study the effects of such systems and stakeholders' views of these tools. Faulconer, Geissler, Majewski, and Trifilo (2014) noted that EWS could help provide positive and negative feedback to students. Students found the positive feedback encouraging and those who were struggling felt like their instructors care about their success. In some situations, teachers might feel like these programs are not useful and feel that they have a better understanding of their students (Soland, 2014). However, human bias can affect judgment and early-alert systems can find hidden patterns about academic success that may not be inherently obvious. In an online environment, it may be difficult to know which students need support. For this reason, researchers integrated an EWS into the launch of a comprehensive academic support service for online students (Britto & Rush, 2013). When students did not log onto the learning management system for more than 72 hours, their advisor received a notification to intervene.

EWS using EDM/LA techniques

When using EDM/LA techniques, EWS utilized past data to make predictions about students' success, sometimes in conjunction with information reported by instructors or student actions. In a study by Belfanz, Herzog, and Mac Iver (2007), researchers used data from students in various Philadelphia schools to see what factors predicted which students would drop out. They found that using four simple factors: poor attendance, poor behavioral grade and failing math or English could correctly identify

60% of high school dropouts. Macfadyen and Dawson (2010) were able to use EDM to create an EWS for educators using LMS data. By examining the data related to how often students access the LMS and their participation in an online class, the authors were able to identify struggling students early so teachers could intervene. Similarly, based on data from an online tutoring system, Casey and Azcona (2017) were able to predict poor performance among students with 85% accuracy for the purposes of early intervention. Moreover, students' choices may also help improve the predictive nature of EWS. Waddington et al. (2016) found that the type of resources students access could help predict their final grade. They noted that this information could be helpful to academic advisors monitoring student progress. With the use of LA techniques, de Freitas et al. (2015) were able to identify struggling students, which allowed academic support professionals and instructor to provide support. These studies demonstrated how EDM/LA techniques can be powerful for improving EWS so that instructors and academic support professionals know which students to target and which students are succeeding on their own. As such, these systems help professionals manage their time and resources better.

Academic Advising and Counseling

Academic advisors and counselors are types of academic support professionals who often oversee the overall academic progress of their students (Huber & Miller, 2013). The benefits of academic advising are numerous and well documented. In a study conducted by Vianden and Barlow (2015), the authors found a strong relationship between the perceived quality of academic advisement and the perceived quality of student services overall. An additional relationship between quality advisement and

institutional loyalty was established. Since students that are more loyal are less likely to leave an institution, this suggests that academic advisement plays an important role in student retention. This corroborates findings from Clay, Rowland, and Packard (2008) who found that intrusive advising, where students are required to meet with an advisor, helped improve retention rates. Similarly, Beck and Mulligan (2014) found that advising effectiveness was a primary and secondary factor in institutional commitment. They found that advising effectiveness had a relationship with institutional commitment itself but had a relationship with degree commitment and academic integrity that in turn had a relationship with institutional commitment. Thompson and Prieto (2013) found that students' satisfaction with academic advising related to higher levels of university satisfaction. These studies relate to the student involvement piece of Lei et al. (2017). Finally, quality advisement enabled students to have a better understanding of their degree requirements allowing them to navigate their educational programs more easily (Schroeder & Terras, 2015; Smith & Allen, 2014).

Unlike predictive analysis for individual classes that might be useful to instructors, overall success models would be of more interest to support professionals outside the classroom. As with EWS, a model created from EDM/LA to predict students' academic success would be valuable to academic counselors and advisors who monitor students' progress. According to the national academic advising association, NACADA's 2011 survey in which academic advisor reported that in their role their work most commonly includes "course scheduling, course registration, and help[ing] students develop a plan of study" (Huber & Miller, 2013). Furthermore, advisors must help "students determine the number of credit hours they can realistically attempt each term."

(Huber & Miller, 2013). Therefore, as the professionals overseeing students' plans of study and helping them choose a credit load, the results of this study could be integral to their work. Moreover, since academic advisors are central to student retention, a tool that can help identify struggling students early could help them in their effort to retain these students and keep them in good academic standing. Mohamed and Waguih (2018) demonstrated how academic advisors used results from EDM analysis of admissions data to help students choose the major which would most likely result in academic success, leading to an increase in retention. While it is important to help students achieve academic success, it is also important to help students pursue their goals.

Summary

The literature on EDM and LA lay the groundwork for this study. These new fields already have a rich history in using data to predict student outcomes accurately in a way that can help improve the work of advisors and other support professionals. In addition, data visualization and EWS represent practical and technical ways to transfer results from EDM and LA methods to those who can use it most. With the use of data visualization, vast amounts of data uncovered in EDM can help advisors digest the information for effective decision-making. EWS provides a streamlines way of alerting these professionals that students are struggling academically. Finally, it is also important to consider findings from the literature that supports how and why predicting academic performance will be beneficial to the work that advisors do.

CHAPTER THREE: METHODOLOGY

The goal of this study was to determine which demographics and prior academic performance measures help to identify students at risk of performing poorly. With the use of a variety of EDM and LA techniques, while using admissions data variables, I attempted to create a functional model for academic support professionals and from this model extract specific factors that have a strong relationship to academic success (Delen, 2010; Şen, Uçar, & Delen, 2012). In this chapter, I detail the specific setting of this study, the participants, and methods for data collection and analysis to answer the following research questions:

1. Using data collected at the point of admission, which predictive algorithm generates the best academic success prediction results on the training data set?
2. What key predictors variables are identified by the best predictive model?

Setting

The setting of this study was Westchester Community College (WCC), part of the State University of New York (SUNY). Like many other community colleges, WCC attracts many students from diverse backgrounds and studying in unique ways. The college offers around 70 courses online for the Spring, Fall, and Summer semesters, most with multiple class sections (WCC, n.d.-a). Additionally, the winter session is comprised of online courses exclusively (WCC, n.d.-d). Currently, 18 degrees can be completed at least 50% online (WCC, n.d.-c). Moreover, many of the students are non-traditional as defined by Hittepole (2017). Individuals 25 or older make up 31% of the college's

student population (WCC, 2017a). Over 69% of students identified as having a minority background and the largest ethnic groups were Hispanic (39.4%), White (30.8%), and black (21%) (WCC, 2017a).

The school employs 14 full- and part-time counselors who provide academic advising (WCC, n.d.-b) to the general student population of 12,571 (WCC, 2017b). The student to counselor ratio is 911:1 but there is no way for a counselor to determine the individualized risk level for new students and provide extra support to those most at risk and make the best use of their time.

Participants

There were no participants in the traditional sense of the term. I did not ask students to answer questions via a survey instrument or interview. However, the data of students enrolling for the first time at WCC since Fall 2015 is of interest. Moreover, the focus of this study was to examine the data of full-time degree-seeking students. Through my data cleaning process, I identified 10,918 unique cases of full-time degree-seeking students entering WCC in this time frame. Due to some missing data, I removed some cases through listwise deletion; I describe the data processing in more detail in Chapter 4. The cleaning process left me with 10,830 cases for analysis, 8114 for training and 2713 for testing.

Data Collection/Cleaning

The academic and demographic data of students pertinent to this study were sex, ethnicity/race, socioeconomic status, high school information, parents' educational history, residency, age, major, grades, credit load, and test scores. These variables have been shown to be effective predictors of academic persistence and achievement as

demonstrated in Table 1 (Abu Tair & El-Halees, 2012; Agnihotri & Ott, 2014; Casanova et al., 2018; Delen, 2010; Marquez-Vera et al., 2013; Pal, 2012; Saheed et al., 2018; Yasmin, 2013; Zimmermann et al., 2015). Since the data I analyzed already existed, I received IRB approval for the study from Boise State University and WCC under exempt status. After the IRB approval from both institutions, I began extracting data for the different variables and compiled it based on the student's anonymous ID number. As soon as practically possible, I replaced these students' ID numbers with random numbers to ensure student privacy. I conducted all processes that include identifying information or actual student ID numbers on a local computer at the college. This ensured that the data remained as protected as it would in usual business processes for the institution.

The approaches to EDM research from Ranjan and Khalil (2008) and Lei et al. (2017) served as a guide for the data cleaning and analysis process. Naturally, the first step in each model is to consider the context or environment in which the research sits and to form hypotheses about how this research can affect that environment. As discussed in chapter one, higher education has been diversifying. In order for academic support professionals, such as academic counselors, to allocate their time best, they must be able to identify which students are most at risk of failing. The central inquiry of this research was: is it possible to predict student success using data collected during students' application process. With these predictions, academic support professionals would be able to categorize students by risk and target intervention methods.

After hypothesis formation, Ranjan and Khalil (2008) recommended researchers collect and pre-process the data for a better understanding of it when building a model. Similarly, Lei et al. (2017) included extracting the raw data and pre-processing it in their

model. I gathered the data through a series of queries of a student information system (specifically PeopleSoft) and link different data elements with the "vlookup" function of Microsoft Excel. This allowed me to build a data set containing information on each variable for each student.

After the data has been pre-processed, Ranjan and Khalil (2008) recommended preparing the data, examining the completeness of the set. Based on how much data and what items were missing, I determined if it is possible to ignore the missing values or if it is necessary to complete further steps to complete the data set (Abu Tair & El-Halees, 2012). I describe my efforts to complete the data set which included listwise deletion and data imputation in more detail in Chapter 4. Additionally, Lei et al. (2017) noted data might need modification. In some cases, I needed to modify the data to make it more uniform (as some outputs represent the same information but in different formats) or to improve the output of the model. For instance, one item of interest was test scores that came from different sources such as SAT scores, ACT scores, and entrance exam scores.

The college uses benchmarks from all three tests to determine students' college-level readiness in English, Math, and Reading. I had to normalize these scores to compare them properly. Additionally, categorical and ordinal variables needed to be dummy coded before analysis. Further, students' GPAs needed modification to create a target variable that takes into account both earning passing grades in individual classes and completing all credits in a semester.

Since the goal was to understand what relationships between admissions data and success in 12 credits or more, I recalculated students' GPAs in a manner that treated all non-passing grades as failures to get an accurate picture of students' academic success.

For instance, a student may take 15 credits but withdraw from 12 of them and receive an "A" in the final three credits. A traditional GPA would show this student as having a 4.0, but this student is no closer to graduating than a student who failed 12 credits and received an "A" in the final three credits.

In these steps of pre-processing and preparing the data, I examined the data for potentially useful trends (Macfadyen & Dawson, 2010; Yasmin, 2013). Through descriptive analysis, I highlighted information on distributions, frequencies, and group differences with data visualizations in the form of charts and graphs (Goldsmith, 1984; Hansen, 1999; Tufte, 1983). Not only did this allow for a baseline understanding of the data set, but it also provided an opportunity to determine if certain data values need to be reduced, combined, or transformed. For instance, some racial/ethnic groups had much fewer cases than others. For analysis, it was more effective to combine these different groups by high and low risk rather than individual groups (Delen, 2010; Waddington et al., 2016).

Data Analysis

At this point, I moved onto the data-mining step described by Lei et al. (2017) and Ranjan and Khalil (2008). To answer the first research question, I ran decision tree, binary logistic regression, and neural network analyses on the training data set to create a model output file in SPSS. In order to run these analyses, I converted students recalculated GPAs (rGPA) into two groups based on the definition of academic success: low risk ($rGPA \geq 2.0$) and high risk ($rGPA < 2.0$).

As its name suggests, binary logistic regression requires a binary target variable (Hatcher, 2013). Deolekar and Abraham (2018) noted that the dependent variable for

decision tree analysis could be either continuous or categorical. When using decision trees to determine dropout rates, Yasmin (2013) used a dichotomous dependent variable. Similar to decision trees, the dependent variable for neural networks can be either continuous (Ramesh et al., 2013; Singh & Kumar, 2013) or categorical (Delen, 2010). So, for all three models, I was able to use the data in the same format.

For the creation of the model, I used data from Fall 2015 to Fall 2017. However, I reserved data from Spring, Summer, and Fall 2018 to test the model for its predictive ability (Alabi, Issa, & Afolayan, 2013; Şen et al., 2012). Then, I compared how the models predicted cases for this new data to the actual student results from those semesters. The model considered the most accurate was the one with the most correct predictions of students' risk levels as compared with actual results (Abu Tair & El-Halees, 2012; Saheed et al., 2018; Singh & Kumar, 2013). This method of comparing actual results to the predicted results is called a confusion matrix (Delen, 2010; Singh & Kumar, 2013). I also considered and noted which categories the models are best at predicting. One model had an overall lower accuracy but better predict at-risk students than another model. In that case, this model might be a more useful EWS tool to academic advisors than one which more accurately predicts students not at-risk (Rogers et al., 2014; Yadav et al., 2012).

Once the best model became apparent with results from the confusion matrix, I used this model to answer the second research question. Feature importance analysis can explain the relationships between the input and output variables used in the model (Alabi et al., 2013). Feature importance analysis helps to determine how important each individual predictor variable is to the accuracy of the model. As Alabi et al. (2013) stated,

"the importance of an independent variable is a measure of how much the network's model-predicted value changes for different values of the independent variable. Moreover, the normalized importance is simply the importance values divided by the largest importance values and expressed as percentages" (p. 26). A variable is more important to a model if a change in its value cause a large change in the predicted value of a given case.

Summary

Educational data mining and learning analytics techniques have the potential to predict student outcomes relating to persistence and achievement. Although none of the studies that I have reviewed examined the same data I collected to predict the same outcomes, this previous literature demonstrated there is potential. Regression, decision trees, and neural networks were promising techniques for this study. It is important to keep in mind that many researchers have used more than one technique as well. Several studies have compared different techniques to gain an understanding of which is the best predictor (Abu Tair & El-Halees, 2012; Saheed et al., 2018; Singh & Kumar, 2013). The results of this study provide a foundation as to how these data variables relate to predicting student success. Even if none of these techniques produce acceptable results, there are more advanced techniques to consider for future research. Others have used ensemble models which integrate two or more techniques to predict academic success (Adejo & Connolly, 2018; Agnihotri & Ott, 2014; Delen, 2010) while others have employed more advanced EDM techniques (Thai-Nghe, Drumond, Krohn-Grimberghe, & Schmidt-Thieme, 2010; Xing et al., 2015).

CHAPTER 4: RESULTS

The goal of this study was to create a predictive model of academic success based on data collected at the point of admissions and from there which data elements were more related to academic success. To achieve this goal, I collected data on degree-seeking students entering a community college over the course of three years and analyzed this data with three common EDM/LA techniques: logistic regression, decision trees, and neural networks. Once the most predictive model was established, feature importance analysis allowed me to determine which factors were most related to academic success within that model. The purpose of these efforts was to establish a model that academic support professionals could use to determine which students are at higher risk for struggling academically. With this information, academic support professionals could target support interventions to these students on an individualized (based on who is predicted to be at risk) or population-wide (based on which factors are most related to academic success) level.

In this chapter, I present the descriptive statistics about the data, efforts to complete the data set, and results from the analyses conducted to answer the research questions. As expected, there were examples of missing data that I needed to address before attempting to analyze the data. This process brought to light some noteworthy findings which I highlight below and elaborate on in the discussion chapter. Once the data set was pre-processed, I was able to run logistic regression, decision tree, and neural network analysis on the variables to determine which approach created the best model.

For each analysis, I made several attempts to improve the results by reprocessing the data in order to build the best possible model with that approach to answer the first research question: which predictive algorithm generates the best academic success predictions result on the training data set? Once the best approach was selected, I ran a feature importance analysis on the selected model to answer the second research question that states: what key predictor variables are identified by the best predictive model?

Training Data

As described in chapter three, methodology, the cases for analysis were limited to full-time degree-seeking students in their first semester at WCC. Using the grade results for students in individual classes I was able to determine which students were enrolled full time. Data from students who were not enrolled in at least 12 semester hours were eliminated from the data set. I cross-referenced the list of full-time students with other data points of age, race/ethnicity, sex, parents' educational history, socioeconomic status, standardized test score, degree program, and high school GPA, to build a more complete data set. I also eliminated data from students who were not pursuing a degree. Finally, I eliminated data from students who were not in their first semester by comparing the admit term to the enrolled term.

Descriptive Statistics

For the training data set, there were 8148 unique cases that met the criteria. However, some cases had missing values in the categories of age, residency, parents' educational background, socioeconomic status, test scores, and high school GPA (Table 3).

Table 3 Number and Percent of Missing Values for the Training Data Set

Variable	N		
	Valid	Missing	Percent missing
Starting semester	8148	0	0%
Residency	8122	26	0.31%
Age	8147	1	0.01%
Prior college credits	8148	0	0%
Academic program	8148	0	0%
Sex	8148	0	0%
Race/Ethnicity	8148	0	0%
Parents' educational history	8145	3	0.04%
Socioeconomic status	8145	3	0.04%
English placement	8138	10	0.12%
Math placement	8138	10	0.12%
Reading placement	8138	10	0.12%
High school GPA	5064	3084	37.84%

Since the number of missing values was relatively small for most of these metrics, cases with missing values were eliminated using listwise deletion which means that the entire record was excluded from the analysis when any single variable was missing (Abu Tair & El-Halees, 2012; Singh & Kaur, 2016). This reduced the number of cases to 8114. However, there was a much higher number of cases missing for high school GPA. Since other researchers have found that a student's GPA in previous course works is important

to predicting student outcomes, this metric was seemed vital to preserve (Delen, 2010; Marquez-Vera et al., 2013; Pal, 2012; Zimmerman et al., 2015). Additionally, the data was not missing at random according to Little's Missing Completely at Random test (Li, 2013). High school GPA was more often missing for older students, so deleting these cases though listwise deletion would disproportionately affect data on older students. To address missing values for this variable, I created a linear regression model based on variables from the valid cases to predict and impute the data for the missing cases (Doreswamy, Gad, & Manjunatha, 2017). The variables I used to create this linear regression model were: starting semester, residency, age, prior college credits, academic program, sex, race/ethnicity, parent's educational history, socioeconomic status, and placement level in math, English, and reading.

There were some findings of note in the linear regression model from the valid cases. Apart from residency and parents' educational history, all variables included were significant factors to the model at $p < .001$. Parents' educational history was significant at $p < .05$, while residency was not significant (Table 4). According to the R^2, statistics the independent variables account for 29% of the variability of the high school GPA. These findings are promising as they suggest that the variables included in this project do have a relationship to academic success, at least at the high school level.

Table 4 Predictive Linear Regression Model to Impute Missing High School GPA Values

Model		Unstandardized coefficients		Standardized coefficients	t	Sig.
		B	Std. Error	Beta		
	(Constant)	76.688	0.828		92.668	0.000
	Starting semester	1.019	0.266	0.047	3.835	0.000
	Residency	0.333	0.21	0.019	1.589	0.112
	Age	-0.244	0.034	-0.088	-7.069	0.000
	Prior college credits	3.187	0.23	0.177	13.83	0.000
	Academic program	0.326	0.09	0.043	3.618	0.000
	Sex	-1.901	0.161	-0.141	-11.77	0.000
	Race/Ethnicity	0.473	0.054	0.11	8.731	0.000
	Parents' educational history	0.35	0.163	0.026	2.145	0.032
	Socioeconomic status	-1.141	0.173	-0.084	-6.582	0.000
	English placement	0.748	0.21	0.049	3.568	0.000
	Math placement	4.326	0.174	0.318	24.899	0.000
	Reading placement	1.961	0.191	0.143	10.244	0.000

Using this model, I imputed a high school GPA for the cases which were missing this piece of data. Tables 5 and 6 contain information on the distribution of the variables in the completed training data set. Table 5 focuses on the continuous variables and

contains information about the range, variances, and measures of central tendency. Table 6 focuses on the categorical variables and contains information on the number and percentage of each value with qualitative descriptions of the possible values

Table 5 Continuous Variables Range, Variance, and Points of Central Tendency in the Training Data Set

	Age (in years)	High school GPA (100-point scale)
Minimum	16	55
Maximum	68	102
Mode	18	76
Median	18	78.57
Mean	19.75	78.57
Variance	21.63	35.35

Table 6 Distribution of Categorical Variables in Training Data Set across Various Values

Categorical inputs	Levels	Description	Number	Percent
Starting semester	0	Started in the Spring or Summer Semester	997	12.3%
	1	Started in a Fall Semester	7117	87.7%
Residency	0	Not a Westchester Resident	1768	21.8%
	1	Westchester Resident	6346	78.2%
Prior college credits	0	Does not have transfer credit	6461	79.6%
	1	Has transfer credit	1653	20.4%
Academic program	0	School of Art, Humanities, and Social Science	2680	33.0%
	1	School of Math, Science, and Engineering	2926	36.1%
	2	School of Business and Professional Careers	2095	25.8%
	3	School of Health Careers, Technology, and Applied Learning	413	5.1%
Sex	0	Female	3839	47.3%
	1	Male	4275	52.7%
Race/Ethnicity	0	Native American	34	0.4%
	1	Asian/Pacific Islander	328	4.0%
	2	Hispanic	1556	19.2%
	3	Black	2091	25.8%
	4	Multiethnic	1580	19.5%
	5	Not Specified	264	3.3%
	6	White	2261	27.9%
Parents' educational history	0	Not a first-generation college student	4492	55.4%
	1	First-generation college student	3622	44.6%
Socioeconomic status	0	Not economically disadvantaged	3528	43.5%
	1	Economically disadvantaged	4586	56.5%
English placement	0	Not college English ready	2299	28.3%
	1	College English ready	5815	71.7%
Math placement	0	Not college math ready	3420	42.1%
	1	College math ready	4694	57.9%
Reading placement	0	Not college reading ready	3234	39.9%
	1	College reading ready	4880	60.1%

Binary Logistic Regression

With the completed data set, I began conducting model building analysis on the training data starting with a binary logistic regression. I began the analysis with the raw data which was compiled from the student information system. Initially, the prediction threshold when determining accuracy was .5. That is, if the prediction threshold that a case would be considered 'at-risk" was .5 or above, then the case was predicted to be "at-risk" if the prediction threshold was below .5 that case was predicted as "not at-risk." This model had an overall accuracy (or percent of correctly identified cases) of 66.4% with a not at-risk accuracy of 76.9% and an at-risk accuracy of 52.6% as shown in Table 7.

Table 7 Confusion Matrix for Binary Logistic Regression – with Raw Data

			Predicted		
			Not at-risk	At-risk	
			0	1	Percent correct
Observed	Not at-risk	0	3531	1030	76.9%
	At-risk	1	1669	1854	52.6%
			Overall percent		66.4%

Although normal distribution is not an assumption of logistic regression (Hatcher, 2013), I examined if any of the scale data was skewed. High school GPA was normally distributed which is clear when looking at a distribution chart of the data overlaid with a normal distribution curve in Figure 4.1 (Ghasemi & Zahediasi, 2012; Hatcher, 2013).

Age was highly skewed with a skewness of 4.295, which is also apparent when looking at a distribution graph of the data in figure 4.2. To address the skewness of the

age data, I attempted to normalize it with a logarithmic transformation (McHugh, Lenz, Reardon, & Peterson; 2012). This reduced the skewness but that metric remained high at 2.974 and the distribution is shown in figure 4.3.

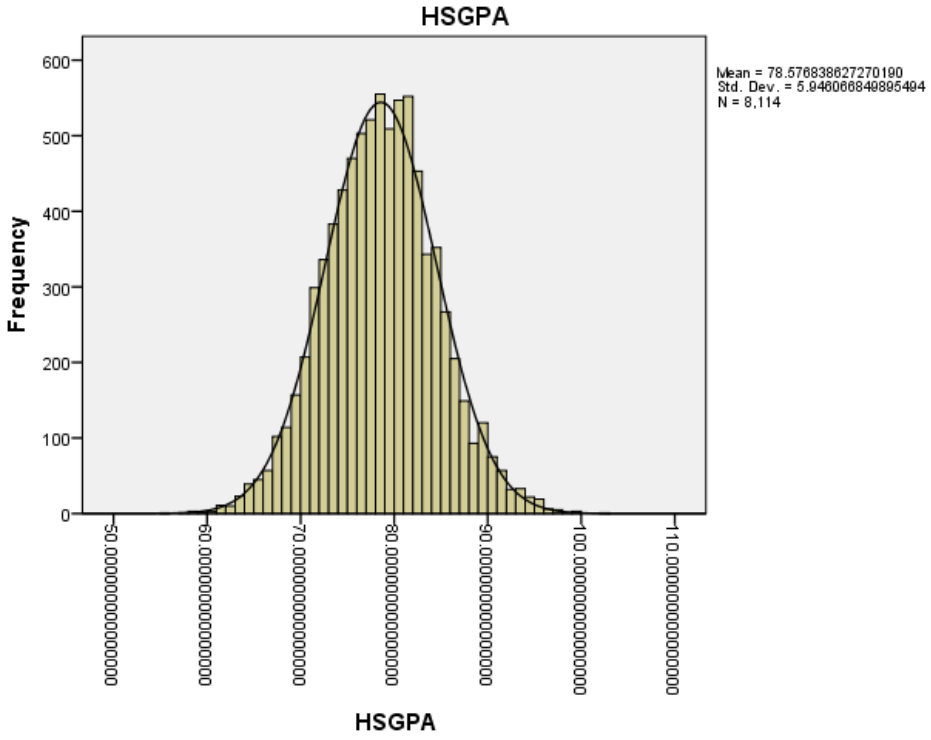

Figure 4.1 Training data set high school GPA distribution

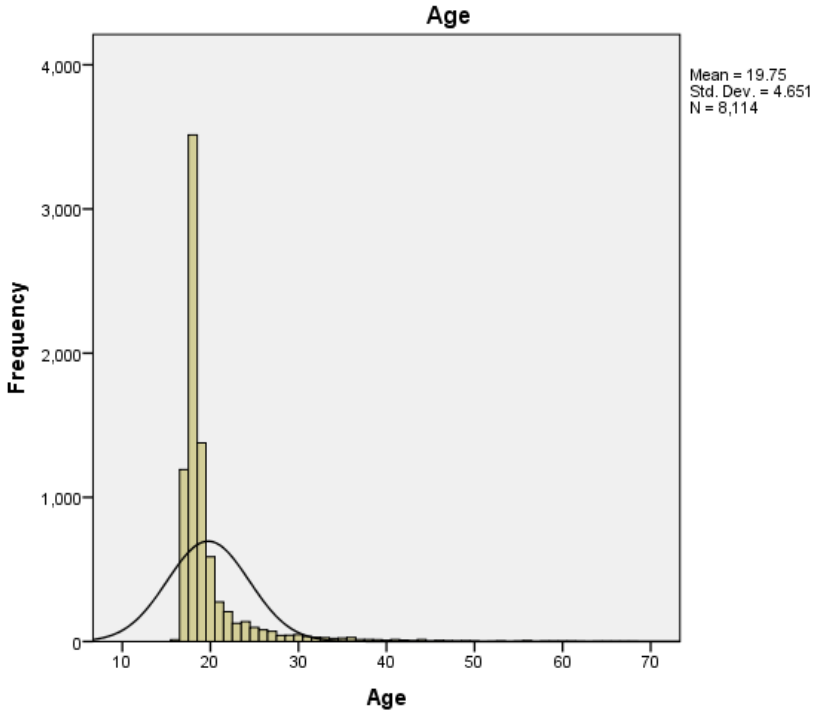

Figure 4.2 Training data set age distribution

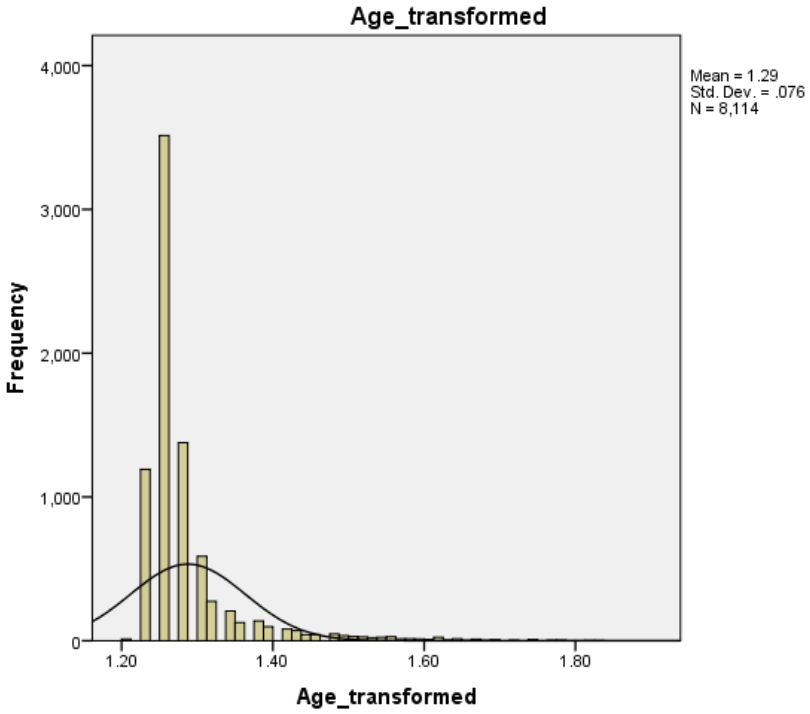

Figure 4.3 Training data set age distribution after logarithmic transformation

With this change, I reran the logistic regression. The overall accuracy did not change, but this model was slightly better at predicting not at-risk students and slightly worse at predicting at-risk students as shown in Table 8.

Table 8 Confusion Matrix for Binary Logistic Regression – with Age Transformed

			Predicted		
			Not at-risk	At-risk	
			0	1	Percent correct
Observed	Not At-risk	0	3540	1051	77.1%
	at-risk	1	1676	1845	52.4%
			Overall percent		66.4%

In another attempt to improve the model's accuracy, I tried to reduce the number of values for the categorical variables, specifically race/ethnicity and academic program code as these were the only two non-dichotomous variables in the model. For each variable, I examined which values showed a larger number of at-risk students than others and reclassified each variable as higher (1) and lower risk (0). This model did not show an improvement in the overall accuracy and further shifted the model towards better predicting not at-risk students at the expense of correct at-risk predictions as shown in Table 9.

Table 9 Confusion Matrix for Binary Logistic Regression – with Reduced Variable Values

			Predicted		
			Not at-risk	At-risk	
			0	1	Percent correct
Observed	Not at-risk	0	3558	1033	77.5%
	At-risk	1	1692	1831	52.0%
			Overall percent		66.4%

The academic program codes represent meta-majors rather than specific degree plans. Therefore, students with the same academic program may be in degree plans with highly different levels of difficulty. I tested if using a dichotomous variable of higher or lower risk degree plans may improve the model. As with the reduction of values for race/ethnicity and academic program, I examined which degree plans had a higher number of students falling into the at-risk category and classified each plan as either higher (1) and lower risk (0). This produced a more accurate model using this technique in terms of overall accuracy, and the accuracy of at-risk and with a slight reduction in not at-risk prediction as shown in Table 10.

Table 10 Confusion Matrix for Binary Logistic Regression – with Academic Plan instead of Program

			Predicted		
			Not at-risk	At-risk	
			0	1	Percent correct
Observed	Not at-risk	0	3552	1039	77.4%
	At-risk	1	1670	1853	52.6%
			Overall percent		66.6%

Finally, I ran a logistic regression using the same variables as the previous model and the two-terms interactions between each variable as predictors. This produced the most accurate model overall and in each classification category as shown in Table 11.

Table 11 Confusion Matrix for Binary Logistic Regression – with Variable Interactions

			Predicted		
			Not at-risk	At-risk	
			0	1	Percent correct
Observed	Not at-risk	0	3588	1003	78.2%
	At-risk	1	1667	1856	52.7%
			Overall percent		67.1%

In table 12, there is a comparison of the overall accuracy and the accuracy of each logistic regression model in the at-risk and not at-risk categories.

Table 12 Comparison of At-risk, Not At-risk, and Overall Accuracy for each Binary Logistic Regression Model

		Raw data	Transformed age	Reduced variable values	Academic plan instead of the program	Variable interactions
Percent correct	Not at-risk	77.1%	77.1%	77.5%	77.4%	**78.2%**
	At-risk	52.4%	52.4%	52.0%	52.6%	**52.7%**
	Overall	66.4%	66.4%	66.4%	66.6%	**67.1%**

Note: highest values in each category in bold.

Due to the highest number of correctly identified cases using the final model, I chose this model as the one with which to test the reserve data. However, since correctly predicting students who are at-risk is of greater importance than overall accuracy in a project like this, I examined the classifications of the model with various thresholds beyond the standard .5 in an attempt to find a balance between precision and recall. The precision measure is the number of true positives, cases which were predicted to be positive and were actually positive, divided by all cases which were predicted to be positive (Saxena, 2018). In this study, a true positive would be an at-risk student who was correctly predicted to be at-risk. The recall measure is the number of true positives divided by the number of positive cases overall (Saxena, 2018).

For the regression model, the balance between precision (predicted at-risk cases were, in fact, at-risk) and recall (at-risk cases not being incorrectly classified as not at-risk), at which both values were approximately 59%, is at the prediction threshold of .46. The various precision, recall, and accuracy levels of the different thresholds highlighted in Figure 4.4.

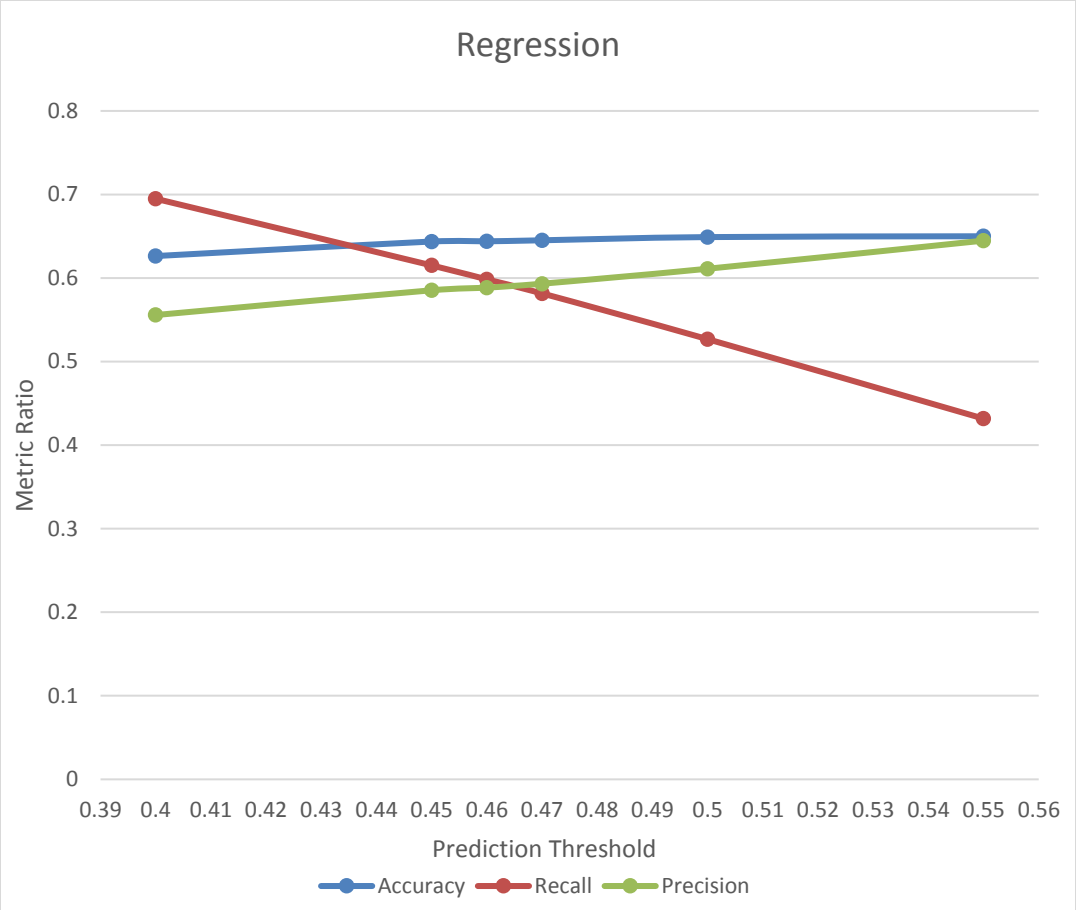

Figure 4.4 Accuracy, Recall, and Precision rates of the final binary logistic regression model at different thresholds

Decision Tree

Once I had selected the best logistic regression model and determined the optimal prediction threshold, I was able to start decision tree analysis. Using the original variables, the decision tree analysis produced an overall accuracy of 66.6%. However, the at-risk prediction was below 50% which represents an accuracy lower than chance as shown in Table 13.

Table 13 Confusion Matrix for Decision Tree Analysis – with Raw Data

			Predicted		
			Not at-risk	At-risk	
			0	1	Percent correct
Observed	Not at-risk	0	3742	849	81.5%
	At-risk	1	1863	1660	47.1%
			Overall percent		66.6%

There were fewer data manipulation options to improve the decision tree analysis due to the nature of this data mining technique which categorizes cases based on overall trends and values (Baradwaj & Pal, 2011; Pal, 2012). The abnormal distribution of age would not affect a decision tree model as it would classify certain ages values as higher or lower risk based on the cases analyzed regardless of distribution. Similarly, recategorizing race or program would not affect the decision tree, as the tree makes those decisions as a function of its analysis. Indeed, an analysis of the variables with these changes produced identical results to the model with the original variables as shown in Table 14.

Table 14 Confusion Matrix for Decision Tree Analysis – with Age Transformed and Reduced Variable Values

			Predicted		
			Not at-risk	At-risk	
			0	1	Percent correct
Observed	Not at-risk	0	3742	849	81.5%
	At-risk	1	1863	1660	47.1%
			Overall percent		66.6%

However, introducing the risk level of the plan into the decision tree analysis did alter the model as this was new information about each case. With this model, the overall accuracy was improved to 66.8% and the accuracy of at-risk prediction improved to 52.9% as shown in table 15. Due to the improved number of classified cases in this model, I chose this as the one to test on the reserve data. See Appendix A for a full visual representation of this decision tree model.

Table 15 Confusion Matrix for Decision Tree Analysis – with Academic Plan instead of Program

			Predicted		
			Not at-risk	At-risk	
			0	1	Percent correct
Observed	Not at-risk	0	3558	1003	77.5%
	At-risk	1	1599	1864	52.9%
			Overall percent		66.8%

As with the logistic regression model, I examined different prediction thresholds to further attempt to improve the classification of this model. For the decision tree model the balance between precision and recall, at which both values were approximately 61%, is at the prediction threshold of .4262. The different levels of precision, recall, and accuracy is highlighted below in Figure 4.5.

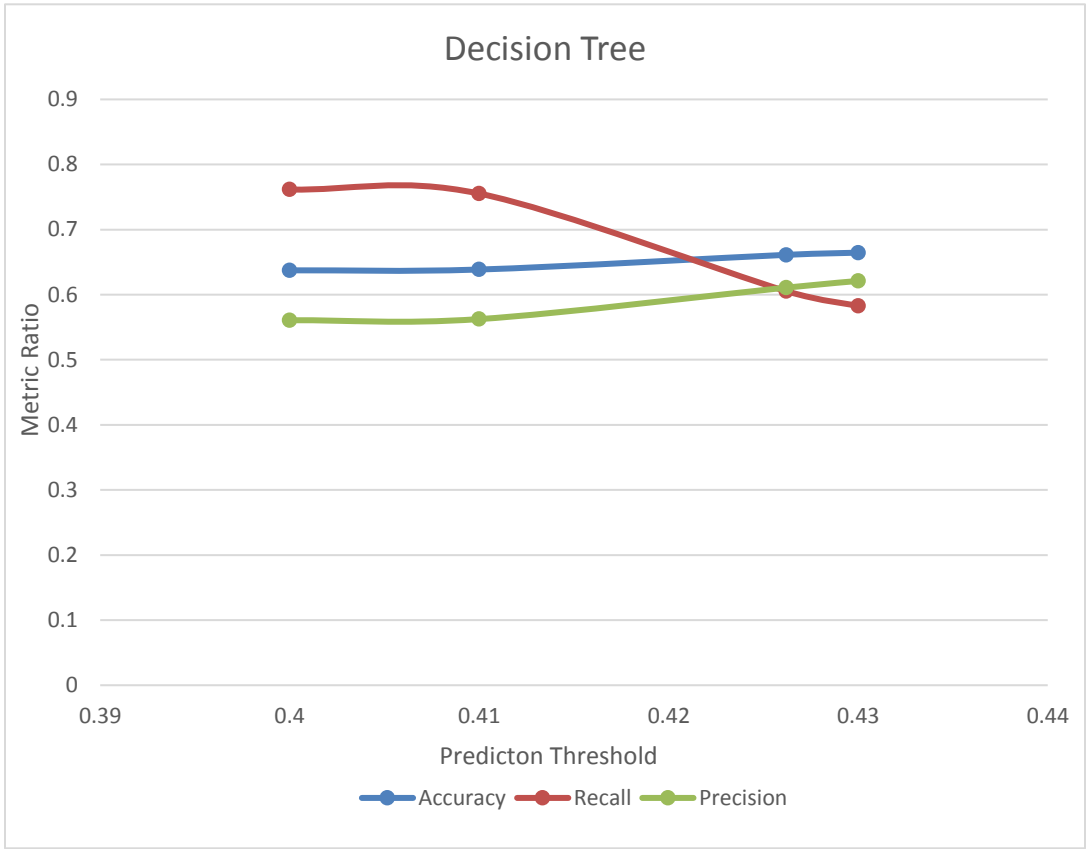

Figure 4.5 Accuracy, Recall, and Precision rates of the final decision tree model at different thresholds

Neural Networks

The final data mining method I attempted to use to build an at-risk model was neural networks. Since the neural network first uses some data to create an analysis of interactions and then validated those interactions (Alabi et al., 2013; IBM, n.d.-b; Zimmerman et al. 2015), I used the validation percentages to determine the best neural

network model. With the original data set, the overall accuracy of the model was 66.6% with a 55.9% accurate prediction of at-risk cases as shown in Table 16.

Table 16 Confusion Matrix for Neural Network Analysis – with Raw Data

			Predicted		
			Not at-risk	At-risk	
			0	1	Percent correct
Observed	Not at-risk	0	1063	365	74.4%
	At-risk	1	464	588	55.9%
			Overall percent		66.6%

As with the logistic regression model, using the transformed age variable did improve the neural network model. This showed an increase in the overall accuracy (67%) and the not at-risk accuracy (75.1%) at a slight expense to the at-risk accuracy (55.8%) as shown in Table 17.

Table 17 Confusion Matrix for Neural Network Analysis – with Age Transformed

			Predicted		
			Not at-risk	At-risk	
			0	1	Percent correct
Observed	Not at-risk	0	1030	341	75.1%
	At-risk	1	444	561	55.8%
			Overall percent		67.0%

I reran the model with the reduced categories for race and program. This showed a slight increase in the overall accuracy of the model (67.1%), a jump in the accuracy of the at-risk prediction (59.5%) and decrease in the not at-risk prediction (73.3%) as shown in Table 18.

Table 18 Confusion Matrix for Neural Network Analysis – with Reduced Variable Values

			Predicted		
			Not at-risk	At-risk	
			0	1	Percent correct
Observed	Not at-risk	0	978	357	73.3%
	At-risk	1	440	647	59.5%
			Overall percent		67.1%

Finally, I reran the analysis with the data relating to academic plans. However, this showed a decrease in the overall accuracy of the model and a decrease in the accuracy of at-risk prediction to below 50% as shown in Table 19.

Table 19 Confusion Matrix for Neural Network Analysis – with Academic Plan instead of Program

			Predicted		
			Not at-risk	At-risk	
			0	1	Percent correct
Observed	Not at-risk	0	1018	296	77.5%
	At-risk	1	520	501	49.1%
			Overall percent		65.1%

In Table 20, there is a comparison of the overall accuracy and the accuracy of each model in the at-risk and not at-risk categories. Due to the higher overall and at-risk accuracy of the model, I choose the second to last neural network model as the test model for the reserved data and for comparison against the selected regression and decision tree models. See Appendix C for a visual representation of this neural network model.

Table 20 Comparison of At-risk, Not At-risk, and Overall Accuracy for each Neural Network Model

		Raw data	Transformed age	Reduced variable values	Academic plan instead of the program
Percent correct	Not at-risk	74.40%	75.10%	73.30%	**77.50%**
	At-risk	55.90%	55.80%	**59.50%**	49.10%
	Overall	66.60%	67.00%	**67.10%**	65.10%

Note: highest values in each category in bold.

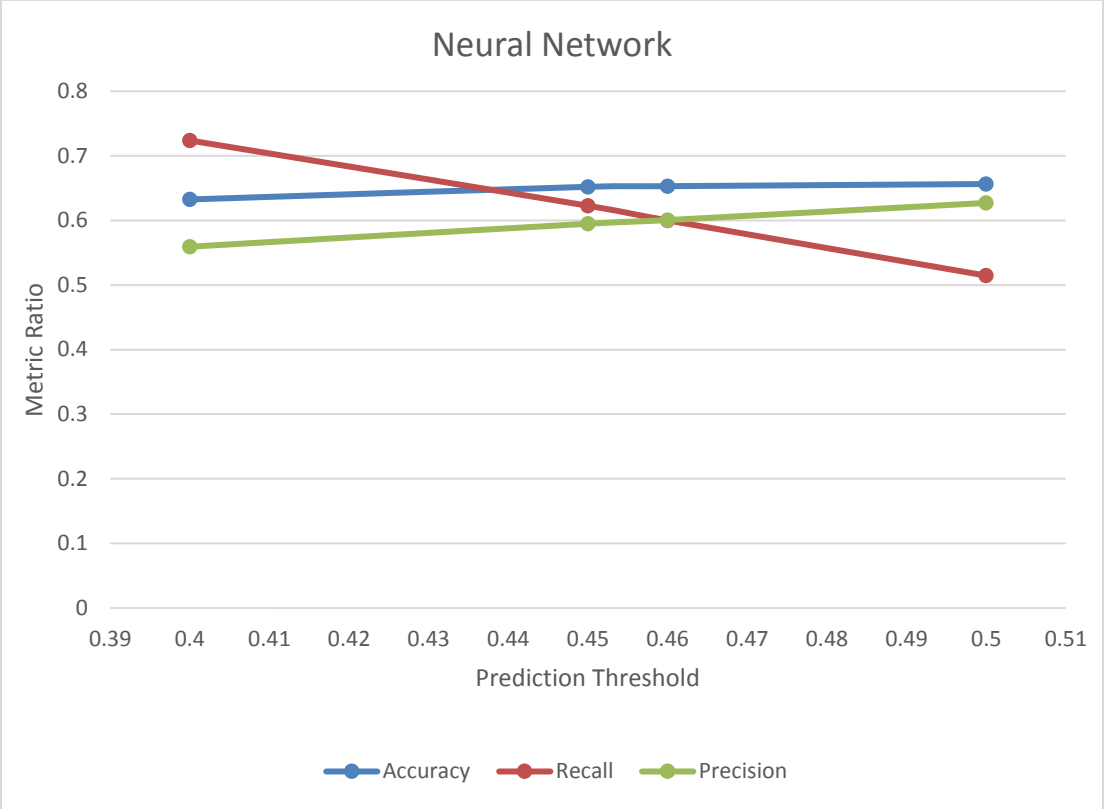

Figure 4.6 Accuracy, Recall, and Precision rates of the final neural network model at different thresholds

For the neural network model, the balance between precision and recall, at which both values were approximately 60%, is at the prediction threshold of .46. You can see the percentages of the recall, precision, accuracy statistics for various thresholds in Figure 4.6.

Testing Data

With an optimal cut off for each model, the next step was to test the model on the reserve data set. I created the testing data set in the same manner as and alongside the training data set. However, before I could test the data set, I needed to ensure that the variables in the training data set matched those for the various models (IBM, n.d.-a). This required the recreation of any new variables to match those in the models. Additionally, I

needed to address any missing variables in the data set in a similar manner to the training data set for congruency.

Descriptive Statistics

As with the training data set, with the elimination of students who did not attend full-time, were not in their first semester, or were not pursuing a degree there were 2770 unique cases for analysis. However, some cases had missing values in the categories of age, residency, parents' educational background, socioeconomic status, test scores, and high school GPA (Table 21).

As with the training data set, there were some variables with a small number of missing values. I eliminated the missing values for residency, parent's educational history, socioeconomic status, and placement testing through listwise deletion. This reduced the overall number of test cases from 2770 to 2716. However, the percentage of missing high school GPA cases was still high, albeit lower than in the training data set. To impute the missing values of high school GPA, I used the same linear regression model created based on the training data. Additionally, like the training data, age was highly skewed with a skewness statistic of 4.165 (Figure 4.7) and remained so even after logarithmic transformation, skewness statistic of 2.971 (Figure 4.8). High school GPA was normally distributed as apparent from Figure 4.9.

Table 21 Number and Percent of Missing Values for Testing Data Set

Variable	N		
	Valid	Missing	Percent missing
Starting semester	2770	0	0%
Residency	2757	13	0.47%
Age	2770	0	0%
Prior college credits	2770	0	0%
Academic program	2770	0	0%
Sex	2770	0	0%
Race/Ethnicity	2770	0	0%
Parents' educational history	2769	1	0.04%
Socioeconomic status	2769	1	0.04%
English placement	2728	42	1.52%
Math placement	2730	40	1.44%
Reading placement	2728	42	1.52%
High school GPA	2033	737	26.61%

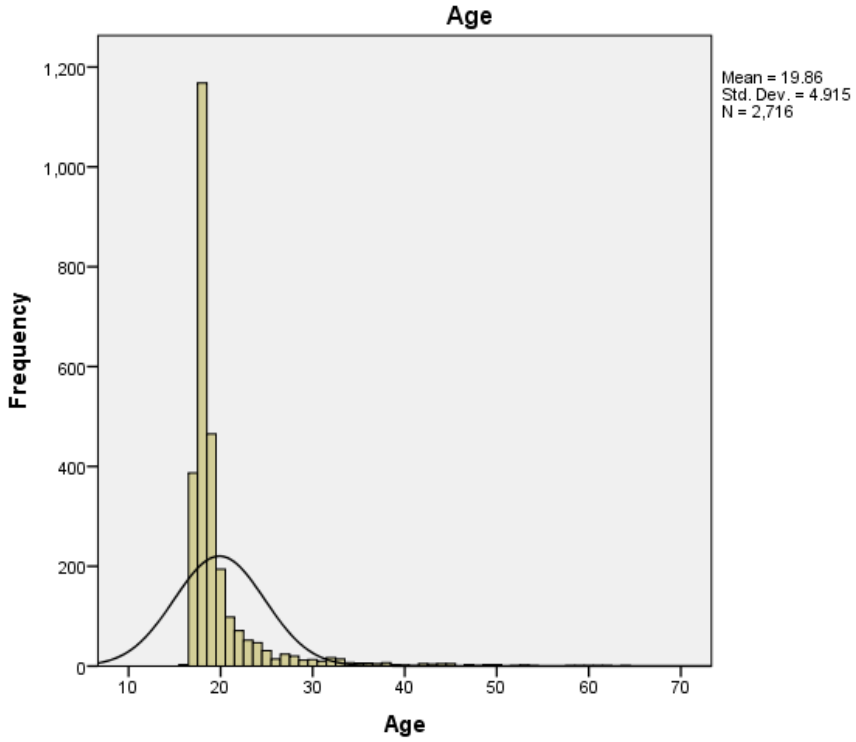

Figure 4.7 Testing data set age distribution

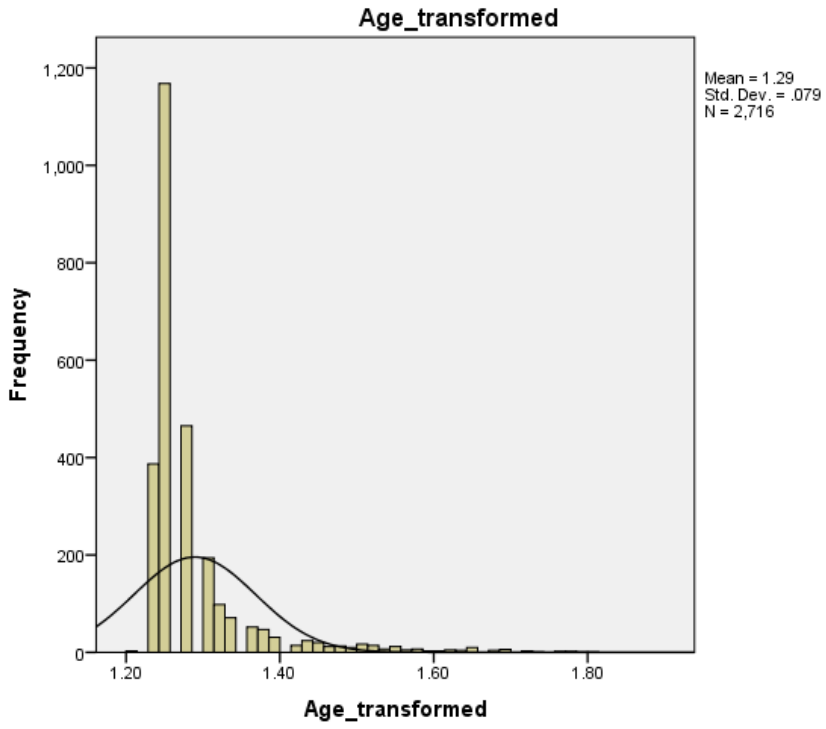

Figure 4.8 Testing data set age distribution after logarithmic transformation

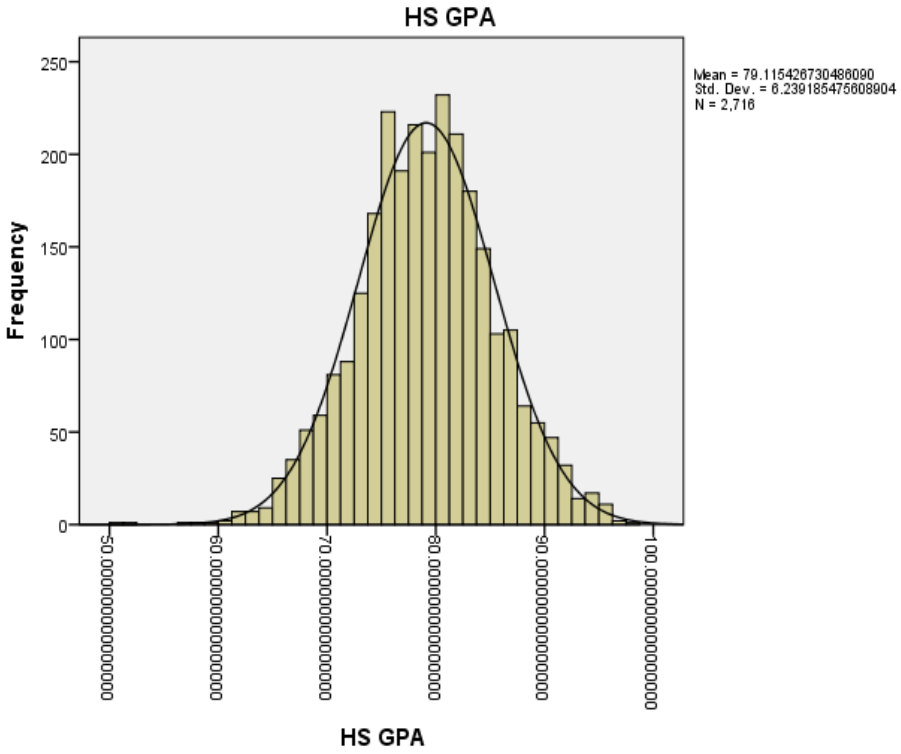

Figure 4.9 Testing data set high school GPA distribution

Table 22 Distribution of Categorical Variables in Testing Data Set across Various Values

Categorical inputs	Levels	Description	Number	Percent
Starting semester	0	Started in the Spring or Summer	511	18.8%
	1	Started in a Fall Semester	2205	81.2%
Residency	0	Not a Westchester Resident	560	20.6%
	1	Westchester Resident	2156	79.4%
Prior college credits	0	Does not have transfer credit	2112	77.8%
	1	Has transfer credit	604	22.2%
Academic program	0	School of Art, Humanities, and Social Science	746	27.5%
	1	School of Math, Science, and Engineering	1049	38.6%
	2	School of Business and Professional Careers	657	24.2%
	3	School of Health Careers, Technology, and Applied Learning	264	9.7%
Sex	0	Female	1248	45.9%
	1	Male	1468	54.1%
Race/Ethnicity	0	Native American	3	0.1%
	1	Asian/Pacific Islander	96	3.5%
	2	Hispanic	528	19.4%
	3	Black	673	24.8%
	4	Multiethnic	609	22.4%
	5	Not Specified	87	3.2%
	6	White	720	26.5%
Parents' educational history	0	Not a first-generation college student	1014	37.3%
	1	First-generation college student	1702	62.7%
Socioeconomic status	0	Not economically disadvantaged	1249	46.0%
	1	Economically disadvantaged	1467	54.0%
English placement	0	Not college English ready	536	19.7%
	1	College English ready	2180	80.3%
Math placement	0	Not college math ready	854	31.4%
	1	College math ready	1862	68.6%
Reading placement	0	Not college reading ready	901	33.2%
	1	College reading ready	1815	66.8%

The descriptive statistics of the categorical and continuous variables are shown in Tables 22 and 23 respectively. Table 22 contains information about the distribution of categorical statistics across various values, while table 23 contains information on the range, variance, and points of central tendency for the continuous variables.

Table 23 Continuous Variables Range, Variance, and Points of Central Tendency in the Testing Data Set

	Age (in years)	High school GPA (100-point scale)
Minimum	16	51
Maximum	64	98
Mode	18	80
Median	18	79.12
Mean	19.86	79.12
Variance	24.16	38.93

Model Testing

The first research question was:

1. Using data collected at the point of admission which predictive algorithm generates the best academic success prediction results on the training data set?

In order to test this hypothesis, I needed to test each model individually and then examine the results from the confusion matrix to identify which model identified the most correct cases overall. The final step necessary before model testing was the recreation of the reduced categorized variables for race, program, and plan using the same criteria for the training data. The data points in the models could be properly matched to the data points in the testing data set (IBM, n.d.-a). Once I had completed this, I had a testing data set that had all the variables used in the models created from the three data mining techniques. I tested the data set with each model and created a

classification table based on the individualized prediction thresholds ascertained from the balance of recall and precision.

Table 24 Comparison of Accuracy, Recall, and Precision rates across Models when Tested on New Data

	Neural Network	Binary Logistic Regression	Decision Tree
Accuracy Percentage	**65.7%**	65.5%	65.6%
Accuracy number of cases	**1784**	1780	1782
Recall	57.4%	55.1%	**57.8%**
Precision	65.7%	**66.3%**	65.5%

Note: highest values in each category in bold

As shown in Table 24, the decision tree had the greatest recall of 57.8% while the regression had the best precision measure of 66.3%. However, the model with the most overall accuracy was the neural network with 65.7% (Table 23). As detailed in chapter 3, the model with the highest number of correctly predicted cases would be considered the best model which was the neural network model. Therefore, I reject the null hypothesis for the research question as one model was better at predicting academic success than the others.

Feature Importance Analysis

The second research question was as follows:

2. What key predictors variables are identified by the best predictive model?

In order to address the second research question, I ran a feature importance analysis on that model using an independent variable importance test (Alabi et al., 2013). As Alabi et al. (2013) stated, “the importance of an independent variable is a measure of

how much the network's model-predicted value changes for different values of the independent variable. Moreover, the normalized importance is simply the importance values divided by the largest importance values and expressed as percentages" (p. 26). Through this test, I found that the high school GPA was the most important factor in the most accurate model. Age and sex also seem to play an important role in this model. Importance statistics for each variable are shown in Table 25 with a graph of these statistics in descending order in Figure 4.10. With these findings, I reject the null hypothesis for research question two as well since there are some factors that were more important to the best prediction model than others.

Table 25 Independent Variable Importance Analysis (Feature Importance Analysis) – Neural Network Model

	Importance	Normalized Importance
Starting semester	0.029	5.4%
Residency	0.012	2.2%
Prior college credits	0.014	2.6%
Sex	0.048	8.8%
Parents' educational history	0.028	5.0%
Socioeconomic status	0.023	4.1%
English placement	0.033	6.0%
Math placement	0.018	3.2%
Reading placement	0.016	2.8%
Race/Ethnicity	0.030	5.5%
Academic program	0.019	3.4%
High school GPA	0.549	100.0%
Age transformed	0.181	33.0%

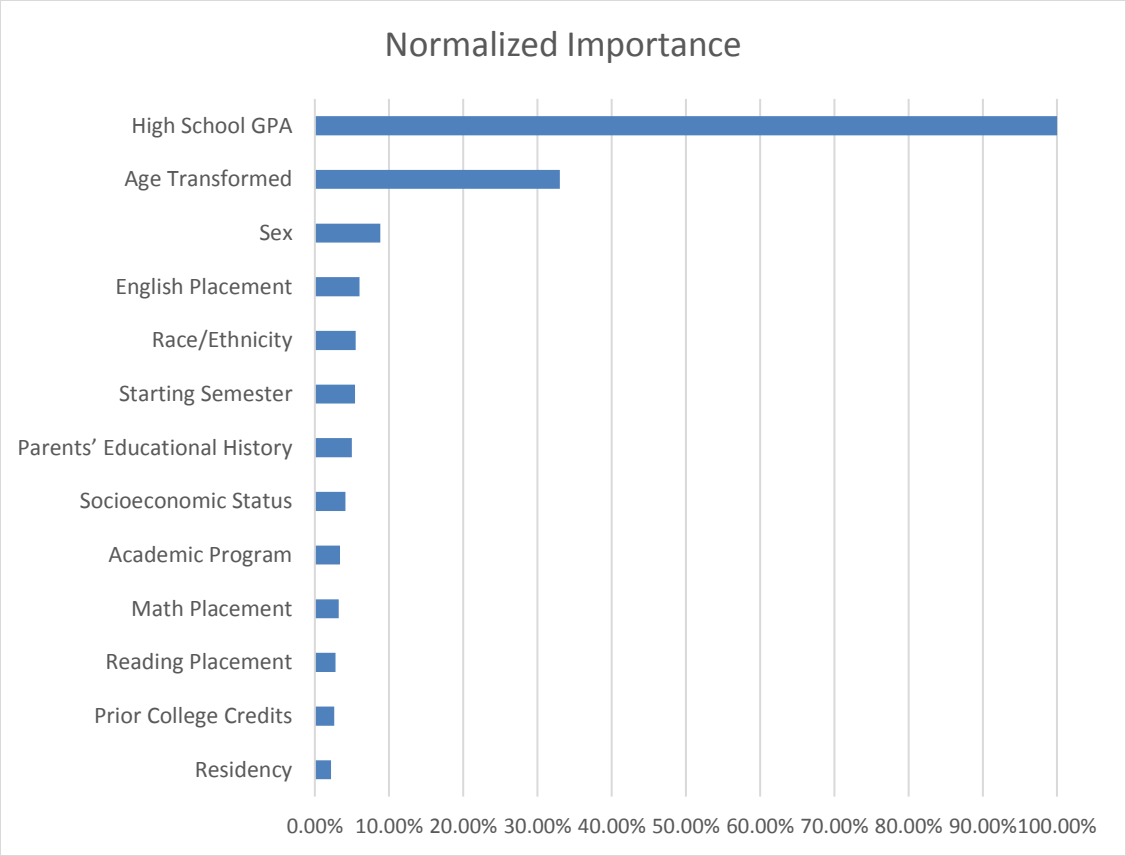

Figure 4.10 Bar graph of normalized importance statistics for neural network model.

Summary

While high school GPA was an important factor in the neural network model, there were several demographic factors significantly related to high school GPA such as race/ethnicity, age, and sex. This implied there are certain populations of students who are more at risk and may be easily identified with this metric. Furthermore, while the neural network was the most accurate predictor of academic success overall, the other models had their strong points such as having a higher recall or precision. Moreover, the visual nature of decision trees can produce a useful categorization tool for educators even if it is not the most accurate overall. Finally, the best model is still only correct about two-thirds of the time. More advanced techniques outside of the scope of this dissertation

might be explored to improve this model. The results of the analysis highlight a myriad of important findings to be further discussed in the final chapter.

CHAPTER 5: DISCUSSION

In this chapter, I begin with a summary of the study's key points including the context, supporting literature, methodology, and results. From there I discuss the findings, how they relate to previous literature, and what aspects of the current study may have affected the results. I also address how the result of this study may have implications for practice and how this research may be utilized in an education setting by academic support professionals. Finally, I touch on recommendations for future research including other approaches and data metrics to consider to potentially improve results.

Summary of the Study

Although the landscape of higher education has changed dramatically, it has not changed evenly. Minority, non-traditional, and online students are more likely to enroll in 2-year public community colleges than 4-year counterparts (National Student Clearinghouse Research Center, 2017b; 2017c). Since 2-year community colleges are often open enrollment institutions and much more affordable, they are accessible to unprivileged groups and as such academic support professionals at these colleges have a unique position to help these disadvantaged individuals succeed. However, this also means that the student body can be made of students with a wide range of academic preparation. Support professionals may have a difficult time distinguishing between students who may struggle and those will be self-sufficient.

I attempted to use information collected from students at the time of admission to predict which students are most in need of support. The most accessible quantitative data

elements such as high school GPA and standardized test scores are not good predictors on their own (Aguinis et al., 2016; Vulperhorst et al., 2018). However, with the use of several data variables collected at the time of admission to the college, I hypothesized it might be possible to build a model to predict academic success and determine which variables are most important to that model. To create this sort model, I used educational data mining and learning analytic techniques. Educational data mining and learning analytics are emerging fields in which researchers use large data set to find underlying relationships between data to help understand learning environments, improve education, and make decisions.

A predictive academic success model can be a powerful early warning systems tool, which could alert support professionals to students who may struggle in college even before classes begin. Aademic support professionals could use this information to target intervention efforts such as study skills workshops, time management techniques, or to just check in on students throughout the semester. This type of intrusive advising and positive student-advisor relationships have been shown to improve student satisfaction and retention rates (Clay et al., 2008; Vianden & Barlow, 2015)

For this study, I analyzed data collected from students by a community college admissions office from Fall 2015 to Fall 2018 and compared it to academic achievement to build a predictive academic success model and identify important factors to that model. The data elements I collected were sex, ethnicity/race, socioeconomic status, parents' educational history high school information, total transfer credits, residency, age, major, grades, credit load, and test scores. I created three models using regression analysis, decision trees, and neural networks, then compared the accuracy of the different models

based on the percentage of correct predictions each model produced and conducted a feature importance analysis to determine the most important factors to the most accurate model.

I examined the records of 10,830 full-time first semester students, 8114 for training and 2713 for testing. The results showed that neural networks were the best model for using admissions data to predict academic success at WCC. Moreover, high school GPA was the most important factor in the neural network model. Other models examined had their strengths both statistically and practically which will be discussed below.

Discussion of Findings

Research Question One

Using data collected at the point of admission which predictive algorithm generates the best academic success prediction results on the testing data set?

By running several iterations of each data analysis technique and modifying the input data, I was able to improve the accuracy of each model. Once I had the most accurate model from the logistic regression, decision tree, and neural network analyses, I adjusted the prediction thresholds to further improve the predictive accuracy. Finally, I compared how well the most accurate versions of each model best predicted cases for the reserve data. The neural network model was accurately predicted the most cases of at-risk versus not at-risk students in the testing data set (Abu Tair & El-Halees, 2012; Saheed et al., 2018; Singh and Kumar, 2013). This model had an overall accuracy of 65.7% followed by the decision tree model (65.6% accurate) and finally the logistic regression model (65.5% accurate).

All three models were better at predicting which students would not be at risk versus predicting students who would be at risk. It is certainly beneficial to know which students will succeed without intervention as this can help optimize resources. However, the preferred error for this study would have been overpredicting the number of cases at risk. As Macfadyen and Dawson (2010) pointed out, it is preferable to provide resources to students who may not ultimately need them than to miss an opportunity to intervene with a student who may not succeed without additional support. Theoretically, if one model had a lower overall accuracy but was better at predicting at-risk students, that model may have had increased practical uses over the others. However, all three models suffered from the issue of underpredicting at-risk students.

Although educational data mining and learning analytics are budding fields, there is already a substantial amount of research on how these methods can be used to predict student performance in individual classes (Baradwaj & Pal, 2011; Yadav et al., 2012; You, 2016). Similarly, other researchers have also used a combination of prior academic history and demographic data to predict dropout rates among undergraduate students before they begin (Agnihotri & Ott, 2014; Delen, 2010; Pal, 2012; Yasmin, 2013). However, it was the aim of this study to connect prior academic history and demographics to academic success. The most accurate model was able to accurately predict students nearly two-thirds of the time but, as noted above, suffered from lower at-risk prediction. Unfortunately, as it stands, the results of this study were not able to bridge the two pieces common in previous literature: using admissions data and predicting academic success. I was not able to predict academic success using admissions

data as accurately as previous studies which employed data from early assignments, attendance, and engagement (Baradwaj & Pal, 2011; Yadav et al., 2012; You, 2016)

Research Question Two

What key predictors variables are identified by the best predictive model?

Since the analysis for the first research question established the neural network model as the most accurate, I used this model to answer the second research question. The results of a feature importance analysis (independent variable importance test) determined that high school GPA was the most important factor to the neural network model. The second most important factor in this model was the transformed variable of age which had a normalized importance value of 33%, meaning it was one-third as important as high school GPA to the model. Sex was the third most important variable with a normalized importance of 8.8%.

These results are supported by literature. In previous studies, some sort of previous GPA (depending on the level of study) was the most common metric found to be an important predictor of academic success at a subsequent institution (Delen, 2010; Marquez-Vera et al., 2013; Pal, 2012; Zimmermann et al., 2015). This follows logically as academic skills are transferable across educational levels. Age was found to be an important factor in a study by Saheed et al. (2018) and sex was found to help predict academic success in studies by Abu Tair and El-Haless (2012), and Casanova et al. (2018).

The most important variable to the neural network also happened to be the one with the most missing values. Since so many of the values for this variable were imputed, it may be a reason for pause as many of the values are not real. Salgado, Azevedo,

Proença, and Vieira (2016) noted that one potential drawback of using a linear regression model to impute missing variables is that the model may overfit the imputed data to the existing data. However, when comparing the training data to the testing data, the percentage of missing high school GPA values decreased from 37.84% to 26.61% while the distribution of this metric remained normal (see Figures 4.1 and 4.9). This suggests that as a higher percentage of high school GPA values are reported to the college, the metric will remain normally distributed which eases the concern of overfitting by the linear regression model.

Comparison to Previous Findings

Previous students have found logistic regression, neural networks and decision trees to be accurate methods for predicting student outcomes measure. Additionally, these methods have been compared against each other as they were in this current study. In this section, I highlight how my findings compare to those of previous researchers.

In several studies, authors have compared the accuracy of different decision tree algorithms. Yadav et al. (2012) used student performance measures, such as quiz scores and attendance, to predict their final grades in a class. They found that the classification and regression trees (CART) method was the most accurate. Similarly, Saheed et al. (2018) found CART and J48 to be equally accurate decision tree models over ID3 when using demographic and academic history measure to predict student performance. J48 and ID3 are different decision tree algorithms. CART and J48 also had the highest precision and recall rates. However, Pal (2012) found ID3 to be more accurate than CART when predicting dropout rates based on demographic and academic history data points.

When using prior academic achievement information, Singh and Kumar (2013) compared the accuracy of several data mining techniques, including decision trees and neural networks, for predicting students' final grades. They found neural networks to be among the most accurate models along with nearest neighbor analysis. Ramesh et al. (2013) had similar findings when predicting academic performance on secondary exit exams based on demographic, social, and academic factors. When comparing several models, including neural networks and decision trees, he found the neural network was the most accurate. Finally, Jishan et al. (2015) found neural networks to be the most accurate algorithm to predict final grades based on academic performance measures in that same class. In this study, naïve bayes were as accurate as the neural network. Furthermore, there was a difference in the recall and precision measures. Though equally predictive, the naïve bayes model had a higher recall measure while the neural network had a higher precision measure. These findings reflect the findings of this study, in that, neural networks were the most or among the most accurate but other models may have better precision or recall measures.

Delen (2010) compared decision trees, neural networks, and logistic regression ability to predict retention rates with the use of demographic and academic data measures. Among these three methods, the decision tree model was the most accurate. However, he also compared these three models to a support vector machine model which was slightly more accurate to the decision tree model. Finally, Shrestha, Orgun, and Busch (2016) used academic performance data to predict if students would accept an offer of admission. Comparing several models, they found that logistic regression was the best model for undergraduate students, and neural networks was the best model for

graduate students. Overall neural networks had a better recall and precision than the other models which also included a decision tree model.

Implications for Practice

Even though the predictive ability of the models created during this study leaves something to be desired, the neural network model was still able to correctly predict a student's risk level two-thirds of the time. Therefore, there is potential usefulness in these models.

As noted, the models overpredicted those not at risk, which means those predicted to be at risk might be acutely at risk. Support professionals, including counselors and advisors, could flag these students for more intrusive interactions with the knowledge that the model missed many of the students who could also benefit from support. Moreover, since all models suffered from this same limitation and had comparable levels of overall accuracy, it might be worth using the decision tree as a rudimentary flow chart for quickly identifying students risk level on a case by case bases. For instance, counselors could quickly reference the decision tree visualization before initial meetings with new students (see Appendix A for decision tree visualization).

Furthermore, high school GPA was found to be the most important factor to the neural network model. It might be worth considering requiring this piece of data from all students due to its predictive ability. Based on the reduced number of missing high school GPA values between the training and testing data sets (the testing data representing more recent students), this may already be a trend. The linear regression analysis highlighted relationships between several other variables and high school GPA. So even in the absence of this data, counselors may examine other data points as proxy variables such as

college course readiness (math, English, and reading) sex, race/ethnicity, and socioeconomic status (Jo et al., 2015).

The connection between high school GPA and the other variables (save residency) is worth examining. Table 26 contains unstandardized and standardized coefficients of the linear regression analysis between high school GPA and the other data points used to impute missing values. From these results, we can see that students from many minority backgrounds (Native American, Black, Hispanic, and Multiethnic) enter WCC with statistically significantly lower high school GPAs than their white peers. Similarly, students with low socioeconomic status are more likely to have a lower high school GPA than their peers from high socioeconomic backgrounds. Moreover, students with lower placement scores, particularly in math, were more likely to have lower high school GPAs. These findings give credibility to the existence of special academic support programs to help students from these backgrounds succeed. These programs include TRIO support services, a federally funded program which provides support to student from low socioeconomic backgrounds, among other groups (U.S. Department of Education, n.d.) and the Educational Opportunity Program, a state-funded program which focuses on helping low-income students (particularly from minority backgrounds) and those with lower placement scores (SUNY, n.d.). Incidentally, both programs operate at WCC.

Table 26 Coefficient from Linear Regression Analysis of High School GPA and other data points

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	79.839	.782		102.126	.000
	Starting Semester	.936	.262	.043	3.574	.000
	Residency	.289	.207	.017	1.392	.164
	Age	-.221	.034	-.080	-6.494	.000
	Prior College Credits	3.089	.228	.171	13.558	.000
	Sex	-1.909	.160	-.142	-11.965	.000
	Parents' Educational History	.345	.161	.026	2.146	.032
	Socioeconomic Status	-1.033	.172	-.076	-6.009	.000
	English Placement	.708	.207	.046	3.425	.001
	Math Placement	4.143	.172	.304	24.082	.000
	Reading Placement	1.913	.189	.139	10.132	.000
	Native American	-5.259	1.254	-.049	-4.194	.000
	Asian/Pacific Islander	2.018	.480	.051	4.202	.000
	Black	-3.135	.257	-.172	-12.220	.000
	Hispanic	-1.617	.220	-.108	-7.339	.000
	Multiethnic	-1.097	.232	-.066	-4.732	.000
	Not Specified	-1.076	.484	-.027	-2.222	.026

	School of Health Careers, Technology, and Applied Learning	.516	.383	.016	1.348	.178
	School of Math, Science, and Engineering	.754	.208	.048	3.627	.000
	School of Business and Professional Careers	-.550	.187	-.039	-2.942	.003

Recommendations for Future Research

To improve the predictive ability of a model like the one at the center of this study, future research may focus on attempting new methods or integrating different data elements into the model. The three analysis techniques I used in this study are relatively common in EDM/LA (Papamitsiou & Economides, 2014) and the data selected were limited to data collected as part of normal business practices.

Other studies to predict students' academic performance or risk level have used an ensemble approach which means that pieces of different models are used in conjunction to improve the predictive ability of a composite model (Adejo & Connolly, 2018; Agnihotri & Ott, 2014). Furthermore, other researchers may consider more advanced EDM/LA techniques. Xing et al. (2015) used a technique called Genetic Programming to predict students' final scores in a class, while Thai-Nghe et al. (2010) made use of a recommender system to also predict student performance.

My intention with this study was to try to predict students' academic performance with the information that a college would already have at their disposal when students start their first semester. The relationship between these metrics and academic performance was well established (see Table 1) but there are other points of data that researchers have shown to be related to academic success. Carnevale and Smith (2018) found that students who work more than 15 hours a week have lower GPAs than those who do not. Similarly, students who are caretakers such as parents of young children can struggle with dedicating time to their studies causing them to have lower academic success outcomes (Wladis, Hachey, & Conway, 2018). Additionally, issues such as housing and food insecurity have been linked to poor academic performance, attendance issues, and the need to delay education (El Zein et al, 2019; Silva et al., 2017). It would not be appropriate to ask for this sort of information on an admissions application, but with the use of an intake form administered by a counselor or case manager, it is possible to collect this data systemically. Combining the new data with application data might create a more robust prediction model.

Although in direct contradiction to the aims of this study, integrating classroom data into a predictive model may improve accuracy. Data related to class behavior or scores on specific assignments have been shown to predict final performance (Baradwaj & Pal, 2011; You, 2016). If such data were collected early enough through some sort of computerized early warning system, this could be integrated with the data used in this study or mentioned in the previous paragraph to help case managers monitor student progress. This sort of model could be especially useful if a reasonable effective preliminary model also existed. For instance, admissions and life circumstances provide

enough data for a predictive base model that early academic performance could enhance with ample time to intervene with the most at-risk students.

Conclusions

The result from the study showed that a neural network model was more accurate than a decision tress or regression model at predicting first-semester academic success among full time students at a community college. Moreover, high school GPA is the most important factor to the neural network model. Even though the research questions were successfully answered, the model still leaves room for improvement. Since there is a dearth of literature related to using the metrics of demographics and prior academic history to academic success, the results from this study provide lessons learned and a jumping-off point for future research.

With the use of additional data points and/or alternative analysis techniques, other researchers may be able to establish a more accurate model for use by academic support professionals. It is imperative that counselors and case managers identify students who may struggle as early as possible to provide the support these students deserve. Socioeconomic status is closely tied to educational level (Berzofsky, Creel, Moore, Smiley-McDonald, & Krebs, 2014) and students from minority groups are more likely to have a lower socioeconomic status (Reeves, Rodigue, & Kneebone, 2016). Not providing necessary support to the diversifying college community can only exacerbate the economic divide that many community college students face.

REFERENCES

Abu Tair, M. M., & El-Halees, A. M. (2012). Mining educational data to improve students' performance: A case study. *International Journal of Information and Communication Technology Research, 2*(2), 140-146. Retrieved from https://pdfs.semanticscholar.org/3711/380a5555ff01e37cb13bde61c8bd95b233f6.pdf

Adejo, O. W., & Connolly, T. (2018). Predicting student academic performance using multi-model heterogeneous ensemble approach. *Journal of Applied Research in Higher Education, 10*(1), 61-75. doi: 10.1108/jarhe-09-2017-0113

Agnihotri, L., & Ott, A. (2014). Building a student at-risk model: An end-to-end perspective. *Proceedings of the 7th International Conference on Educational Data Mining, London, United Kingdom.* 209-212. Retrieved from http://educationaldatamining.org/EDM2014/uploads/procs2014/short%20papers/209_EDM-2014-Short.pdf

Aguinis, H., Culpepper, S. A., & Pierce, C. A. (2016). Differential prediction generalization in college admissions testing. *Journal of Educational Psychology, 108*(7), 1045-1059. doi: 10.1037/edu0000104

Akanmu, S. A. & Jamaluddin, Z. (2016). Designing information visualization for higher education institutions: A pre-design study. *Journal of Information and Communication Technology, 15*(1), 145-163. Retrieved from https://www.researchgate.net/publication/303792392_DESIGNING_INFORMATION_VISUALIZATION_FOR_HIGHER_EDUCATION_INSTITUTIONS_A_PRE-DESIGN_STUDY

Alabi, M. A., Issa, S., & Afolayan, R. B. (2013). An application of artificial intelligent neural networks and discriminant analyses on credit scoring. *Mathematical Theory and Modeling, 3*(11), 20-28. doi: 10.36478/jmmstat.2013.47.54

Astin, A. W. (1999). Student involvement: A developmental theory for higher education. *Journal of College Student Development, 40*(5), 518-529. Retrieved from https://www.middlesex.mass.edu/ace/downloads/astininv.pdf

Astin, A. W. (2012) *Assessment for excellence: The philosophy and practice of assessment and evaluation in higher education.* Lanham, MD: Rowman & Littlefield.

Baradwaj, B. K., & Pal, S. (2011). Mining educational data to analyze students' performance. *International Journal of Advanced Computer Science and Applications, 2*(6), 63-69. Retrieved from https://arxiv.org/ftp/arxiv/papers/1201/1201.3417.pdf

Beck, H. P., & Milligan, M. (2014). Factors influencing the institutional commitment of online students. *Internet and Higher Education, 20,* 54-56. doi: 10.1016/j.iheduc.2013.09.002

Belfanz, R., Herzog, L., & Mac Iver, D. J. (2007). Preventing student disengagement and keeping students on the graduation path in urban middle-grades schools: Early identification and effective interventions. Educational Psychologist, 42(4), 223-235. doi: 10.1080/00461520701621079

Berzofsky, M., Creel, D., Moore, A., Smiley-McDonald, H., & Krebs, C. (2014). *Measuring socioeconomic status (SES) in the NCVS: Background, options, and recommendations* (Report No. 0213170.001.002.001). Retrieved from https://www.bjs.gov/content/pub/pdf/Measuring_SES-Paper_authorship_corrected.pdf

Britto, M., & Rush, S. (2013). Developing and implementing comprehensive student support services for online students. Journal of Asynchronous Learning Networks, 17(1), 29-42. doi: 10.24059/olj.v17i1.313

Calvet Liñán, L., & Pérez, Á. A. J. (2015). Educational data mining and learning analytics: Differences, similarities, and time evolution. *Universities and Knowledge Society Journal, 12*(3), 98-112. doi: 10.7238/rusc.v1213.2515

Carnevale, A. P., & Smith, N. (2018). *Balancing work and learning: Implications for low-income students.* Retrieved from Georgetown University's website: https://1gyhoq479ufd3yna29x7ubjn-wpengine.netdna-ssl.com/wp-content/uploads/Low-Income-Working-Learners-FR.pdf

Casanova, J. R., Cervero, A., Núñez, J. C., Almeida, L. S., & Bernardo, A. (2018). Factors that determine the persistence and dropout of university students. *Psicothema, 30*(4), 408-414. doi: 10.7334/psicotherma2018.155

Casey, K., & Azcona, F. (2017) Utilizing student activity patterns to predict performance. *International Journal of Educational Technology in Higher Education, 14*(4). doi: 10.1186/s41239-017-0044-3

Clay, M. N., Rowland, S., & Packard, A. (2008). Improving undergraduate online retention through grated advisement and redundant communication. *Journal of College Student Retention, 10*(1), 93-102. doi: 10.2190/CS.10.1.g

de Freitas, S., Gibson, D., Du Plessis, C., Halloran, P., Williams, E., Ambrose, M., … Arnab, S. (2015). Foundations of dynamic learning analytics: Using university student data to increase retention. *British Journal of Educational Technology, 46*(6), 1175-1188. doi: 10.111/bjet.12212

Deolekar, S., & Abraham, S. (2018). Tree-based classification of tabla strokes. *Current Science, 115*(9), 1724-1731. doi: 10.18520/cs/v115/i9/1724-1731

Delen, D. (2010). A comparative analysis of machine learning techniques for student retention management. *Decision Support Systems, 49,* 498-506. doi: 10.1016/j.dss.2010.06.003

Doreswamy., Gad, I., & Manjunatha, B. R. (2017). *Proceedings from the 2017 International Conference on Advances in Computing, Communications and Informatics, Manipal, India,* 1327-1334. doi: 10.1109/ICACCI.2017.8126025

El Zein, A., Shelnutt, K. P., Colby, S., Vilaro, M. J., Zhou, W., Green, G.,…Mathews, A. E. (2019). Prevalence and correlates of food insecurity among U.S. college students: A multi-institutional study. *BMC Public Health, 19*. doi: 10.1186/s12889-019-6943-6

Faulconer, J., Geissler, J., Majewski, D., & Trifilo, J. (2014). Adoption of an early-alert system to support university student success. *Delta Kappa Gamma Bulletin, 80*(2), 45-48.

Ghasemi, A., & Zahediasi, S. (2012). Normality tests for statistical analysis: A guide for non-statisticians. *International Journal of Endocrinology and Metabolism, 10*(2), 486-489. doi: 10.5812/ijem.3505

Goldsmith, E. (1984). *Research into illustration: An approach and a review.* Cambridge, UK: Cambridge University Press.

Hansen, Y. (1999). Graphic tools for thinking, planning, and problem solving. In R. Jacobsen (Ed.), *Information design* (pp. 193-220). Cambridge, MA: MIT Press.

Hatcher, L. (2013). *Advanced statistics in research: Reading, understanding, and writing up data analysis results.* Saginaw, MI: Shadow Finch Media.

Higher Education Services Corporation. (n.d.). *Full-time study.* Retrieved from https://www.hesc.ny.gov/partner-access/financial-aid-professionals/tap-and-scholarship-resources/tap-coach/48-full-time-study.html

Hittepole, C. (2017). *Nontraditional students: Supporting changing student populations.* Retrieved from https://www.naspa.org/images/uploads/main/Hittepole_NASPA_Memo.pdf

Huber, J., & Miller, M. A. (2013). Implications for advisor job responsibilities at 2-and 4-year institutions. Retrieved from http://www.nacada.ksu.edu/Resources/Clearinghouse/View-Articles/Advisor-Job-Responsibilities.aspx

Huff, D. (1982). *How to Lie with Statistics.* New York, NY: W. W. Norton & Company.

IBM. (n.d.-a). *Building and scoring models.* Retrieved January 21, 2020 from https://www.ibm.com/support/knowledgecenter/en/SSCVKV_10.1.0/SPSSIntegration/Campaign/SPSSMA/How_to_Build_Score_Models.html

IBM. (n.d.-b). *IBM SPSS neural networks 24.* Retrieved January 21, 2020 from ftp://public.dhe.ibm.com/software/analytics/spss/documentation/statistics/24.0/en/client/Manuals/IBM_SPSS_Neural_Network.pdf

Jishan, S. T., Rashu, R. I., Haque, N., & Rahman, R. M. (2015). Improving accuracy of students' final grade prediction model using optimal equal width binning and synthetic minority over-sampling technique. *Decision Analytics, 2*(1). doi: 10.1186/s40165-014-0010-2

Jo, I. H., Kim, D., & Yoon, M. (2015). Constructing Proxy Variables to Measure Adult Learners' Time Management Strategies in LMS. *Educational Technology & Society, 18*(3), 214–225. Retrieved from https://pdfs.semanticscholar.org/a90e/f99aa92bcee605d3f9f3ecfc445af65c56f4.pdf?_ga=2.268885810.239462878.1538795200-1028378358.1530842583

Kennedy-Clark, S. (2013). Research by design: DBR and the higher degree research student. *Journal of Learning Design, 6*(2), 108-122. doi: 10.5204/jld.v6i2.128

Lei, X.-F., Yang, M., & Cai, Y. (2017). Educational data mining for decision-making: A framework based on student development theory. *Advances in Engineering Research, 117*, 628-641. Retrieved from https://download.atlantis-press.com/article/25873946.pdf

Leung, K. C. (2015). Preliminary empirical model of crucial determinants of best practice for peer tutoring on academic achievement. *Journal of Educational Psychology, 107*(2), 558-579. doi: 10.1037/a0037698

Li, C. (2013). Little's test of missing completely at random. *Stata Journal, 13*(4), 795-809. doi: 10.1177/1536867X1301300407

Lin, S., Fortuna, J., Kulkarni, C., Stone, M., & Heer, J. (2013). Selecting semantically-resonant colors for data visualization. *Computer Graphics Forum, 32*(3), 401-410. doi: 10.1111/cgf.12127

Macfadyen, L. P., & Dawson, S. (2010). Mining LMS data to develop an "early warning system" for educators: A proof of concept. *Computer and Education, 54*, 588-599. doi: 10.1016/j.compedu.2009.09.008

Márquez-Vera, C., Romero, C., & Ventura, S. (2013). Predicting school failure using data mining. *The IEEE Journal of Latin-American Learning Technologies, 8*(1), 7-14. doi: 10.1109/rita.2013.2244695

McHugh, E. A., Lenz, J. G., Reardon, R. C., & Peterson, G. W. (2012). The effects of using model-reinforcing video on information-seeking behavior. *Australian Journal of Career Development, 21*(1), 14-21. doi: 10.1177/103841621202100103F

Mohamed, M. H., & Waguih, H. M. (2018). A proposed academic advisor model based on data mining classification techniques. *International Journal of Advanced Computer Research 8*(36), 129-136. doi: 10.19101/IJARC.2018.836003

Monmonier, M. (1996). *How to lie with maps* (Second ed.). Chicago, IL: University of Chicago Press.

National Center for Education Statistics. (2012). *Number and percentage of graduate students taking distance education or online classes and degree programs, by selected characteristics: Selected years, 2003-04 through 2011-12.* Retrieved from https://nces.ed.gov/programs/digest/d15/tables/dt15_311.32.asp

National Center for Education Statistics. (2014a). *Distance learning.* Retrieved from https://nces.ed.gov/fastfacts/display.asp?id=80

National Center for Education Statistics. (2014b). *Total fall enrollment in degree-granting postsecondary institutions, by attendance status, sex, and age: Selected years, 1970 through 2024.* Retrieved from https://nces.ed.gov/programs/digest/d14/tables/dt14_303.40.asp?referrer=report

National Center for Education Statistics. (2016). *Total fall enrollment in degree-granting postsecondary institutions, by level of enrollment, sex, attendance status, and race/ethnicity of student: Selected years, 1976 through 2015.* Retrieved from https://nces.ed.gov/programs/digest/d16/tables/dt16_306.10.asp?current=yes

National Center for Education Statistics. (2018). *The condition of education 2018: Undergraduate enrollment.* Retrieved from https://nces.ed.gov/programs/coe/indicator_cha.asp

National Student Clearinghouse Research Center. (2017a). *Completing college – national– 2017.* Retrieved from https://nscresearchcenter.org/wp-content/uploads/SignatureReport14_Final.pdf

National Student Clearinghouse Research Center. (2017b). *Completing college – national by race and ethnicity – 2017.* Retrieved from https://nscresearchcenter.org/signaturereport12-supplement-2/

National Student Clearinghouse Research Center. (2017c). *Current term enrollment – fall 2017.* Retrieved from https://nscresearchcenter.org/current-term-enrollment-estimates-fall-2017/

Ozaki, C. C. (2016). Possible selves, possible futures: The dynamic influence of changes in the possible selves on community college returnees' persistence decisions. *Journal of College student Retention: Research, Theory & Practice, 17*(4), 413-436. doi: 10.1177/1521025115579248

Pal, S. (2012). Mining educational data to reduce dropout rates of engineering students. *International Journal of Information Engineering and Electronic Business, 2*, 1-7. doi: 10.5815/ijieeb.2012.02.01

Papamitsiou, Z., & Economides, A. A. (2014). Learning analytics and educational data mining in practice: A systematic literature review of empirical evidence. *Educational Technology & Society, 17*(4), 49-64. Retrieved from https://www.jstor.org/stable/jeductechsoci.17.4.49

Pardo, A., Mirriahi, N., Martinez-Maldonado, R., Jovanovic, J., Dawson, S., & Gaešvić, D. (2016). *Proceedings from the 2016 Learning Analytics and Knowledge Conference, Edinburgh, United Kingdom,* 474-478. doi: 10.1145/2883851.2883870

Peña-Ayala, A. (2014). Educational data mining: A survey and a data mining-based analysis of recent works. *Expert Systems with Applications, 41*, 1432-1462. doi: 10.1016/j.eswa.2013.08.042

Prakash, B. R., Hanumanthappa, M., & Kavitha, V. (2014). Big data in educational data mining and learning analytics. *International Journal of Innovative Research in*

Computer and Communication Engineering, 2(12), 7515-7520. doi: 10.15680/ijircce.2014.0212044

Ramesh, V., Parkavi, P., & Ramar, K. (2013). Predicting student performance: A statistical and data mining approach. *International Journal of Computer Application, 63*(8), 35-39. Retrieved from http://www.ece.uvic.ca/~rexlei86/SPP/otherswork/studdmapproach.pdf

Ranjan, J., & Khalil, S. (2008). Conceptual framework of data mining process in management education in India: An institutional perspective. *Information Technology Journal, 7*(1), 16-23. doi: 10.3923/itj.2008.16.23

Reeves, R., Rodrigue, E., & Kneebone, E. (2016). *Five evils: Multidimensional poverty and race in America.* Retrieved from The Brookings Institution's website: https://www.brookings.edu/wp-content/uploads/2016/06/ReevesKneeboneRodrigue_MultidimensionalPoverty_FullPaper.pdf

Rogers, T., Colvin, C., & Chiera, B. (2014). Modest analytics: Using the index method to identify students at risk of failure. *Proceedings from the 2016 Learning Analytics and Knowledge Conference, Indianapolis, IN*, 118-122. doi: 10.1145/2567574.2567617

Romero, C., & Ventura, S. (2010). Educational data mining: A review of the state of the art. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), 40*(6), 601-618. doi: 10.1109/TSMCC.2010.2053532

Romero, C., & Ventura, S. (2013). Data mining in education. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 3*, 12-27. doi: 10.1002/widm.1075

Rose, K. (2017). Data on demand: A model to support the routine use of quantitative data for decision-making in access services. *Journal of Access Services, 14*(4), 171-187. doi: 10.1080/15367967.2017.1394195

Saheed, Y. K., Oladele, T. O., Akanni, A. O., & Ibrahim, W. M. (2018). Student performance prediction based on data mining classifications techniques. *Nigerian Journal of Technology, 37*(4), 1087-1091. doi: 10.4314/njt.v37i4.31

Salgado, C. M., Azevedo, C., Proença, H., & Vieira, S. M. (2015). Missing data. MIT Critical Data, *Secondary analysis of electronic health records* (pp. 143-162). Cambridge, MA: Springer.

Saxena, S. (2018, May 11). Precision vs recall [Web log post]. Retrieved from https://towardsdatascience.com/precision-vs-recall-386cf9f89488

Şen, B., Uçar, E, & Delen, D. (2012). Predicting and analyzing secondary education placement-test scores: A data mining approach. *Expert Systems with Applications, 39*(10), 9468-9476. doi: 10.1016/j.eswa.2012.02.112

Schroeder, S. M., & Terras, K. L. (2015). Advising experiences and needs of online, cohort, and classroom adult graduate learners. *NACADA Journal, 35*(1), 42-55. doi: 10.12930/NACADA-13-044

Shapiro, D., Dundar, A., Huie, F., Wakhungu, P. K., Yuan, X., Nathan, A., & Hwang, Y. (2017, September). *Tracking Transfer: Measures of Effectiveness in Helping Community College Students to Complete Bachelor's Degrees* (Signature Report No. 13). Retrieved from National Student Clearinghouse Research Center website: https://nscresearchcenter.org/signaturereport13/

Shrestha, R. M., Orgun, M. A., & Busch, P. (2016). Offer acceptance prediction of academic placement. *Neural Computing and Applications, 27*, 2351-2368. doi: 10.1007/s00521-015-2085-7

Siemens, G. (2012). Learning analytics: Envisioning a research discipline and a domain of practice. *Proceedings of the 2nd International Conference on Learning Analytics and Knowledge, Vancouver, BC*, 4-8. doi: 10.1145/2330601.2330605

Siemens, G., & Baker, R. S. J. D. (2010). Learning analytics and educational data mining: Towards communication and collaboration. *Proceedings of the 2nd International Conference on Learning Analytics and Knowledge, Vancouver, BC,* 252-254. doi: 10.1145/2330601.2330661

Silva, M. R., Kleinert, W. L., Sheppard, A. V., Cantrell, K., A., Freeman-Coppadge, D. J., Tsoy, E.,…Pearrow, M. (2017). The relationship between food insecurity, housing instability, and school performance among college students in an urban university. *Journal of College Student Retention, 19*(3), 284-299. doi: 10.1177/1521025115621918

Sin, K., & Muthu, L. (2015). Application of big data in education data mining and learning analytics – A literature review. *ICTACT Journal of Soft Computing, 5*(4), 1035-1049. doi: 10.21917/ijsc.2015.0145

Singh, S., & Kumar, V. (2013). Performance analysis of engineering students for recruitment using classification data mining techniques. *International Journal of Computer Science Engineering & Technology, 3*(2), 31-37. Retrieved from https://pdfs.semanticscholar.org/6f2f/e472de38ad0995923668aa8e1abe19084d09.pdf

Singh, W., & Kaur, P. (2016). Comparative analysis of classification techniques for predicting computer engineering student academic' performance. *International Journal of Advanced Research in Computer Science, 7*(6). doi: 10.26483/ijarcs.v7i6.2793

Smith, C. L., & Allen, J. M. (2014). Does contact with advisors predict judgments and attitudes consistent with student success? A multi-institutional study. *NACADA Journal, 34*(1), 50-63. doi: 10.12930/NACADA-13-019

Soland, J. (2014). Is "Moneyball" the next big thing in education? *Kappan Magazine, 96*(4), 64-67. doi: 10.1177/0031721714561450

SUNY. (n.d.). *Educational opportunity program (EOP).* Retrieved January 21, 2020 from https://www.suny.edu/attend/academics/eop/

Thai-Nghe, N., Drumond, L., Krohn-Grimberghe, A., & Schmidt-Thieme, L. (2010). Recommender system for predicting student performance. *Procedia – Computer Science, 1*, 2811-2819. doi: 10.1016/j.procs.2010.08.006

Thompson, L. R., & Prieto, L. C. (2013). Improving retention among college students: Investigating the utilization of virtual advising. *Academy of Educational*

Leadership Journal, 17(4), 13-26. Retrieved from http://www.abacademies.org/articles/aeljvol17no42013.pdf

Tufte, E. R. (1983). *The visual display of quantitative information.* Cheshire, CT: Graphics Press.

Tversky, B. (2001). Spatial schemas in depictions. In M. Gattis (Ed.), *Spatial schemas and abstract thought* (pp. 79-112). Cambridge, MA: MIT Press.

U.S. Department of Education. (n.d.). *Federal TRIO programs – home page.* Retrieved January 21, 2020 from https://www2.ed.gov/about/offices/list/ope/trio/index.html

Vianden, J., & Barlow, P. (2015). Strengthening the bond: Relationships between academic advising quality and undergraduate student loyalty. *NACADA Journal, 35*(2), 15-27. doi: 10.12930/NACADA-15-026

Vulperhorst, J., Lutz, C., de Kleijn, R., & van Tartwijk, J. (2018). Disentangling the predictive validity of high school grades for academic success in university. *Assessment & Evaluation in Higher Education, 43*(3), 399-414, doi: 10.1080/02602938.2017.1353586

Waddington, R. J., Nam, S., Lonn, S., & Teasley, S., D. (2016). Improving early warning systems with categorized course resources usage. *Journal of Learning Analytics, 3*(3), 263-290. doi:10.18608/jla.2016.33.13

Westchester Community College. (n.d.-a). *Course schedules.* Retrieved January 20, 2019, from http://www.sunywcc.edu/academics/courses/

Westchester Community College. (n.d.-b). *Office personnel.* Retrieved January 20, 2019, from http://www.sunywcc.edu/student-services/counseling/counseling-staff/

Westchester Community College. (n.d.-c). *What programs can I take online?* Retrieved January 20, 2019, from http://www.sunywcc.edu/academics/online-education/what-programs-can-i-take-online/

Westchester Community College. (n.d.-d). *Winter session.* Retrieved March 15th, 2018, from http://www.sunywcc.edu/academics/wintersession/

Westchester Community College. (2017a). *Common data set 2017-2018.* Retrieved from http://www.sunywcc.edu/cms/wp-content/uploads/2018/02/2017-2018-Common-Data-Set.pdf

Westchester Community College. (2017b). *ESS final figures. Fall semester 2017: Student profile.* Retrieved from http://www.sunywcc.edu/cms/wp-content/uploads/2016/12/Fall-Student-Profile-2017.pdf

Wladis, C., Hachey, A. C., & Conway, K. (2018). No time for college? An investigation of time poverty and parenthood. *The Journal of Higher Education, 89*(6), 807-831. doi: 10.1080/00221546.2018.1442983

Wibrowski, C. R., Matthews, W. K., & Kitsantas, A. (2017). The role of a skills learning support program on first-generation college students' self-regulation, motivation, and academic achievement: A longitudinal study. *The Journal of College Student Retention: Research, Theory & Practice, 19*(3), 317-332. doi: 10.1177/1521025116629152

Xing, W., Guo, R., Perakovic, E., & Goggins, S. (2015). Participation-based student final performance prediction model through interpretable genetic programming: Integrating learning analytics, educational data mining and theory. *Computers in Human Behavior, 47,* 168-181. doi: 10.1016/j.chb.2014.09.034

Yadav, S. K., Bharadwaj, B., & Pal, S. (2012). Data mining applications: A comparative study for predicting student's performance. *International Journal of Innovative Technology & Creative Engineering, 1*(12), 13-19. Retrieved from https://arxiv.org/ftp/arxiv/papers/1202/1202.4815.pdf

Yasmin, D. (2013). Application of the classification tree model in predicting learner dropout behaviour in open and distance learning. *Distance Education, 34*(2), 218-231. doi: 10.1080/01587919.2013.793642

You, J. W. (2016). Identifying significant indicators using LMS data to predict course achievement in online learning. *Internet and Higher Education, 29,* 23-30. doi: 10.1016/j.jheduc.2

Zimmermann, J., Brodersen, K. H., Heinimann, H. R., & Buhmann, K. M. (2015). A model-based approach to predicting graduate-level performance using indicators of undergraduate-level performance. *Journal of Educational Data Mining*, *7*(3), 151-176. Retrieved from https://jedm.educationaldatamining.org/index.php/JEDM/article/view/JEDM07015.11.003

APPENDIX A

Decision Tree Table and Visualization

Table A.1 Decision Tree Table

Node	0		1		Total		Predicted Category	Parent Node	Primary Independent Variable				
	N	Percent	N	Percent	N	Percent			Variable	Sig.[a]	Chi-Square	df	Split Values
0	4591	56.6%	3523	43.4%	8114	100.0%	0						
1	261	32.2%	550	67.8%	811	10.0%	1	0	High School GPA	.000	815.416	7	<= 71.1100000000000000
2	303	37.4%	508	62.6%	811	10.0%	1	0	High School GPA	.000	815.416	7	(71.1100000000000000, 73.5492040648007400]
3	756	46.6%	868	53.4%	1624	20.0%	1	0	High School GPA	.000	815.416	7	(73.5492040648007400, 77.0600000000000000]

4	444	54.7%	367	45.3%	811	10.0%	0	0	High School GPA	.000	815.416	7	(77.0600000000000000, 78.5720000000000000]
5	976	60.2%	646	39.8%	1622	20.0%	0	0	High School GPA	.000	815.416	7	(78.5720000000000000, 81.6095368427453600]
6	552	68.0%	260	32.0%	812	10.0%	0	0	High School GPA	.000	815.416	7	(81.6095368427453600, 83.4711441992277600]
7	604	75.3%	198	24.7%	802	9.9%	0	0	High School GPA	.000	815.416	7	(83.4711441992277600, 85.9977446405134200]
8	695	84.7%	126	15.3%	821	10.1%	0	0	High School GPA	.000	815.416	7	> 85.9977446405134200

9	109	20.8%	414	79.2%	523	6.4%	1	1	Age	.000	99.316	2	<= 19.0
10	18	32.7%	37	67.3%	55	0.7%	1	1	Age	.000	99.316	2	(19.0, 20.0]
11	134	57.5%	99	42.5%	233	2.9%	0	1	Age	.000	99.316	2	> 20.0
12	178	29.9%	418	70.1%	596	7.3%	1	2	Age	.000	59.294	2	<= 20.0
13	27	45.8%	32	54.2%	59	0.7%	1	2	Age	.000	59.294	2	(20.0, 23.0]
14	98	62.8%	58	37.2%	156	1.9%	0	2	Age	.000	59.294	2	> 23.0
15	489	41.3%	696	58.7%	1185	14.6%	1	3	Age	.000	65.470	2	<= 19.0
16	125	52.1%	115	47.9%	240	3.0%	0	3	Age	.000	65.470	2	(19.0, 23.0]
17	142	71.4%	57	28.6%	199	2.5%	0	3	Age	.000	65.470	2	> 23.0
18	369	58.2%	265	41.8%	634	7.8%	0	4	Plan Risk Level	.000	13.994	1	Lower Risk
19	75	42.4%	102	57.6%	177	2.2%	1	4	Plan Risk Level	.000	13.994	1	Higher Risk
20	881	58.7%	619	41.3%	1500	18.5%	0	5	Age	.000	17.239	1	<= 23.0
21	95	77.9%	27	22.1%	122	1.5%	0	5	Age	.000	17.239	1	> 23.0

22	342	71.8%	134	28.2%	476	5.9%	0	6	Age	.000	21.732	2	<= 18.0
23	166	58.5%	118	41.5%	284	3.5%	0	6	Age	.000	21.732	2	(18.0, 23.0]
24	44	84.6%	8	15.4%	52	0.6%	0	6	Age	.000	21.732	2	> 23.0
25	518	74.0%	182	26.0%	700	8.6%	0	7	English Placement	.024	5.093	1	College English ready
26	86	84.3%	16	15.7%	102	1.3%	0	7	English Placement	.024	5.093	1	Not college English ready
27	467	88.3%	62	11.7%	529	6.5%	0	8	Age	.001	15.060	1	<= 18.0
28	228	78.1%	64	21.9%	292	3.6%	0	8	Age	.001	15.060	1	> 18.0
29	89	24.1%	281	75.9%	370	4.6%	1	9	Plan Risk Level	.005	7.913	1	Lower Risk
30	20	13.1%	133	86.9%	153	1.9%	1	9	Plan Risk Level	.005	7.913	1	Higher Risk
31	86	66.2%	44	33.8%	130	1.6%	0	11	Sex	.003	8.990	1	Female
32	48	46.6%	55	53.4%	103	1.3%	1	11	Sex	.003	8.990	1	Male

33	133	28.0%	342	72.0%	475	5.9%	1	12	Residency	.049	3.888	1	Westchester Resident
34	45	37.2%	76	62.8%	121	1.5%	1	12	Residency	.049	3.888	1	Not a Westchester Resident
35	383	43.5%	497	56.5%	880	10.8%	1	15	Plan Risk Level	.007	7.185	1	Lower Risk
36	106	34.8%	199	65.2%	305	3.8%	1	15	Plan Risk Level	.007	7.185	1	Higher Risk
37	106	57.3%	79	42.7%	185	2.3%	0	16	Plan Risk Level	.003	8.794	1	Lower Risk
38	19	34.5%	36	65.5%	55	0.7%	1	16	Plan Risk Level	.003	8.794	1	Higher Risk
39	110	75.9%	35	24.1%	145	1.8%	0	17	Prior College Credits	.021	5.307	1	Does not have transfer credits

40	32	59.3%	22	40.7%	54	0.7%	0	17	Prior College Credits	.021	5.307	1	Has transfer credits
41	189	63.0%	111	37.0%	300	3.7%	0	18	Sex	.020	5.389	1	Female
42	180	53.9%	154	46.1%	334	4.1%	0	18	Sex	.020	5.389	1	Male
43	711	57.4%	528	42.6%	1239	15.3%	0	20	Prior College Credits	.021	5.341	1	Does not have transfer credits
44	170	65.1%	91	34.9%	261	3.2%	0	20	Prior College Credits	.021	5.341	1	Has transfer credits
45	97	69.3%	43	30.7%	140	1.7%	0	23	Sex	.000	13.347	1	Female
46	69	47.9%	75	52.1%	144	1.8%	1	23	Sex	.000	13.347	1	Male
47	450	72.7%	169	27.3%	619	7.6%	0	25	Math Placement	.030	4.714	1	College math ready

48	68	84.0%	13	16.0%	81	1.0%	0	25	Math Placement	.030	4.714	1	Not college math ready
49	152	83.5%	30	16.5%	182	2.2%	0	28	Parent's Educational History	.004	8.337	1	Not a first-generation college student
50	76	69.1%	34	30.9%	110	1.4%	0	28	Parent's Educational History	.004	8.337	1	First-generation college student

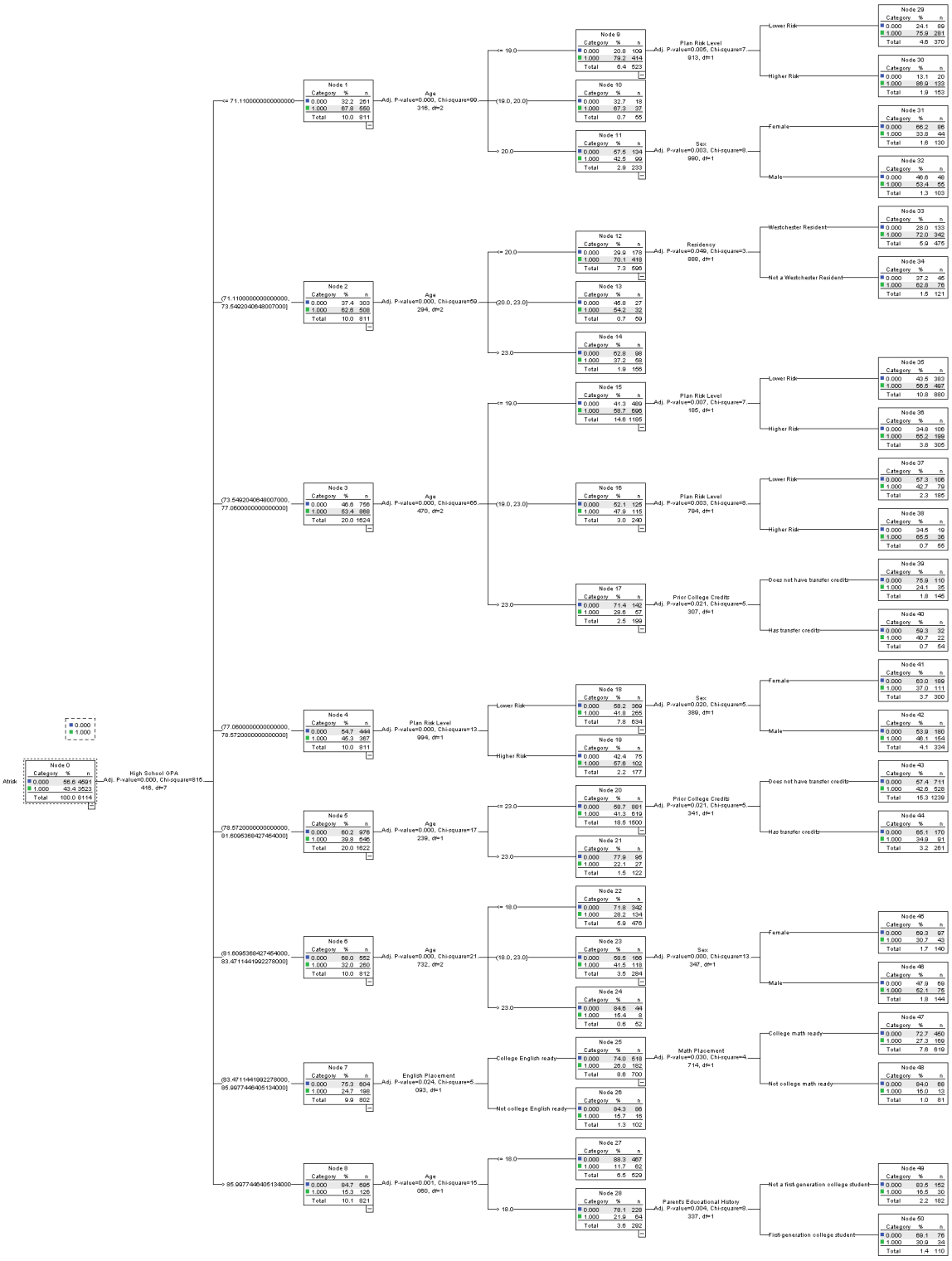

Figure A.1 Visual representation the decision tree model

APPENDIX B

Neural Network Visualization

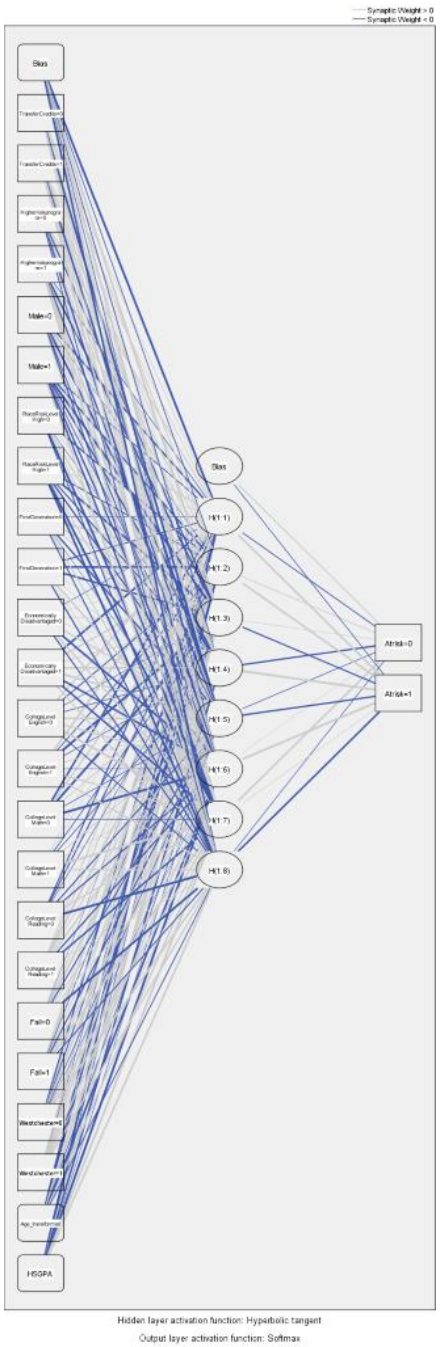

Figure B.1 Visual representation of the neural network model