Higher Education
Quality Council
of Ontario
An agency of the Government of Ontario

Using Predictive Modelling to Inform Early Alert and Intrusive Advising Interventions and Improve Retention

Ross Finnie, Tim Fricker, Eda Bozkurt,
Wayne Poirier, Dejan Pavlic

Published by

The Higher Education Quality Council of Ontario

1 Yonge Street, Suite 2402
Toronto, ON Canada, M5E 1E5

Phone:	(416) 212-3893
Fax:	(416) 212-3899
Web:	www.heqco.ca
E-mail:	info@heqco.ca

Cite this publication in the following format:

Finnie, R., Fricker, T., Bozkurt, E., Poirier, W., Pavlic, D. (2017) *Using Predictive Modelling to Inform Early Alert and Intrusive Advising Interventions and Improve Retention.* Toronto: Higher Education Quality Council of Ontario.

Higher Education
Quality Council
of Ontario

The opinions expressed in this research document are those of the authors and do not necessarily represent the views or official policies of the Higher Education Quality Council of Ontario or other agencies or organizations that may have provided support, financial or otherwise, for this project.

Acknowledgements

This project involved dozens of people who all played important roles and this report would not have been possible without their involvement and support.

First, to Mohawk College staff:

- Gary Jennings, Hetal Patel and Craig Cook who facilitated the Assessments for Success program. This is a fabulous program that has great data and provides valuable support to students.

- Brian Maloney, George Rombes, Helen Sheridan and Shannon Gould from Corporate Reporting and Institutional Research. Each has provided hours upon hours of data management, cleaning and Student Entrance Survey administration.

- The Student Success Advisers and Barb Russell who collectively made an enormous shift in practice to adopt the ClockWork software program and track advising services data.

- The Mohawk Executive Group (past and present) for their ongoing support of this work.

- The additional members of the Mohawk research team who are not listed as authors, but had important operational roles, including Rachel Matthews and Megan Pratt.

Second, to the EPRI team:

- Kaveh Afshar for his insights and for keeping the project moving at the EPRI end while serving as Project Manager at EPRI.

- Michael Dubois, Project Manager at EPRI, for picking up from Kaveh to contribute to the final sets of revisions and for guiding the paper through the final stages of production.

- John Sergeant for his administrative talents, which further helped keep the project on track.

- The rest of the EPRI research team who helped brainstorm and otherwise provide suggestions on the research.

Third, to HEQCO for funding this work and for setting up the Access and Retention Consortium. This collaborative and supportive approach to research has been enormously helpful. At the more general level, EPRI would like to thank HEQCO for its support across a range of projects over the years. EPRI would not be where it is today without that support.

And finally, to the other members of the consortium for their feedback and suggestions on the design of this project.

Executive Summary

Origins and Outline of the Project

In 2012, Mohawk College solicited the support of the Education Policy Research Initiative (EPRI) to collect and use administrative and other data on students held by Mohawk as part of a broad initiative to improve student success based on the principle of evidence-based decision making.

The first project involved analyses to better understand student retention at Mohawk using both descriptive and statistical modelling approaches. This work led to the development of a predictive model to identify students at risk of leaving college early.

In 2015, Mohawk and EPRI applied to and became part of the Higher Education Quality Council of Ontario's (HEQCO) Access and Retention Consortium (ARC) to undertake a project that would build on this earlier work. The purpose was to update, refine and extensively test the predictive model, which would then be used to inform and assess a set of alternative advising interventions put in place for students entering Mohawk College in Fall 2015.

The overall study has two corresponding phases of investigation. Phase 1, which is the subject of this report, represents the work that EPRI undertook to further develop the predictive model of student retention at Mohawk College; to test the predictive model and the predictions it generates; to use this model to generate predictions of the probability that any given incoming student would leave Mohawk early; to use these predictions to divide students into three different risk groups (high, medium, low); to analyze the distribution of students across these different risk groups according to a range of individual, program and other characteristics; and to then calculate participation rates in existing student advising programs across the three risk groups.

Phase 2 (and the second report) will consist of an analysis of the specific advising interventions offered to students of the 2015 entry cohort, including how the effects of these interventions vary across student risk groups as identified by the predictive model.

In addition, this paper not only goes through the development and application of the predictive model and related analysis, but also provides a primer on the use of predictive modelling within the context of predictive analytics and student retention. The motivation for such a primer comes from the wide use of the term "predictive modelling" as pertaining to student retention, and the potential usefulness of identifying the relevant concepts.

The paper thus begins by discussing in general terms what constitutes a predictive modelling approach in the context of postsecondary education (PSE) student retention. It then attempts to explain what this means by describing each step involved in creating the predictive model. The steps include: model specification, model estimation, initial model testing and using the model to generate predicted leaving rates at the

individual level. The subsequent steps involve testing the model further to determine the accuracy of its predictions and, finally, the model is used to divide Mohawk College students into three risk-level groups.

Key Findings

Some of the specific findings are as follows:

- The main determinants of leaving Mohawk College before graduation as identified by the predictive model are as follows: females, students aged 23 and older[1], and Graduate Certificate students have lower leaving rates than others; Certificate students and those with lower incoming grades (especially D plus or lower) leave at higher rates; there are substantial differences in leaving rates across schools; region (urban, rural, international) is not statistically significant; of the Student Entry Survey (SES) based risk categories, only "career clarity" and especially "education commitment" are significant; reading and math assessment scores also matter.
- The individual-level predicted student leaving rates (with leaving defined as students who did not start the second semester of their program) generated by the predictive model, which was estimated over the 2005 to 2012 entry cohorts and then tested for the 2013 and 2014 entry cohorts, closely tracked actual leaving rates.
- The actual leaving rates of students in the three student risk categories defined using the predictive model are 24% for high-risk students, 14% for medium-risk students, and 9% for low-risk students in the 2013–14 test cohorts, further showing how the predictive model effectively differentiates students according to their risk levels.
- Finally, contrary to common perceptions by student affairs practitioners that the students who participate in advising and support programs are those who do not actually need them (Dietsche, 2012), it was found that students in the high- and medium-risk groups participated the most in existing student advising programs at about 22%, while only 14% of the low-risk students participated (all results for the 2013–14 test cohorts).

Implications for Mohawk College Policies

For Mohawk College, these findings have a number of important implications. First, changing the cut-off predicted leaving rates used to define the different student risk categories results in substantially different groups of students of different sizes with different characteristics, which has implications for student support policies. How the predictive model is used thus represents sets of policy decisions rather than being solely a statistical exercise.

Although not all of the SES-based risk categories are statistically significant, "career clarity" and especially "education commitment" are two notable exceptions, and are worth further review to see what they

[1] The reason for choosing 23 as the cut-off for the oldest age group was to have larger sample sizes needed during the training of the predictive model.

capture and how students identified as at-risk by these variables could be helped. Reading and math assessment scores also matter, highlighting the need to further help students become prepared for college in terms of these essential competencies.

The fact that having a low high-school average is a particular risk factor for leaving early (albeit among many other risk factors identified in the modelling approach used here) suggests that the college may want to revisit its minimum and/or conditional requirements in applicable programs to determine if policy changes, including enhanced student support initiatives, could lead to improved retention.

Conclusion and Avenues for Future Research

Overall, the findings suggest that employing administrative and related student-level data to develop and then use predictive models of student retention represents a promising practice to assist institutions in better understanding student retention, in targeting students at higher risk of leaving with student support programs, and in testing and further developing policies, programs and services that could have a positive effect on student persistence, retention and graduation rates.

One avenue for further research is, therefore, for more institutions to develop and employ the same sort of predictive models of student retention based on the experiences of their own students. This would provide a more complete picture of student retention, allow for better targeting of student programs, and lead to more solid statistical testing of the effects of student support initiatives put in place.

At the same time, the models employed could be improved by:

1. Adding additional variables based on data already held by institutions, such as the need for financial aid, or program choice data from college applications;
2. Adding further variables by linking to other data sources — such as adding socio-economic information on a student by using postal code information to connect to census-based information on the student’s background neighbourhood;
3. Also adding to the models “early reporting” on students such as attendance, early assessments or grades, which reflect student behaviour and outcomes after their initial point of entry into the institution; and
4. Diving into the potentially massive sources of electronic information increasingly being collected on students, such as those related to course participation and engagement, including online interaction with learning management systems.

Finally, consideration should be given to developing and testing other kinds of predictive models, such as those recently proposed based on advanced machine-learning algorithms to see if they lead to better predictions and, if so, if any such improvements outweigh their complexity in both their development and application.

Table of Contents

List of figures

List of tables

1. Introduction

1.1 The Meaning, Development and Use of Predictive Models

In the most general sense, a predictive model can be thought of as a mathematical function that produces predictions for an outcome of interest given the values for predictor variables. In the context of postsecondary education (PSE) and student retention, the aim of such models can be to predict whether a student will leave the institution before graduating based on the observable characteristics of the student at a given time. Such predictions can be used to target students with interventions and supports based on their risk level.

In this context, there are a range of approaches to predicting students who are at risk of leaving early. At one end of the spectrum, simple risk indicators such as a student's gender, incoming grades, and hours of paid work can be used as crude univariate indicators of "high-risk" students based on simple descriptive analyses undertaken for other students at other institutions (even in other countries), which identify general relationships between leaving rates and these student attributes.

For example, one simple approach is to classify a student as being at high risk of leaving if he or she has a certain number of these risk factors (characteristics). However, this approach relies on very simple statistical analytical approaches (i.e., the univariate indicators described) of entirely different populations of students.[2]

A somewhat improved version of this approach is to use past data and analysis based on the students/institution in question to establish the empirical relationships underlying the indicators developed. However, the result is still a set of crude and, to a significant degree, arbitrary zero-one (univariate) indicators, which are again used in an ad hoc fashion to identify students at risk.

At the other end of this spectrum is the development of regression-type or more sophisticated models (including some of the advanced machine-learning algorithms) to arrive at the model which best predicts student behaviour (e.g., leaving early) at the relevant institution.

The modern predictive modelling literature is characterized by a few key features and advantages. First, models are built from the ground up with the aim of maximizing predictive accuracy in new data. To achieve this, a portion of the available data is used as a training sample and once the model is developed, its predictive accuracy is assessed using the testing sample. This is the key difference between, for example, a descriptive regression model and a predictive regression model.

[2] A particularly unfortunate (and common) example of the extreme errors such an approach can generate in the Canadian context is to focus on "first generation students" — who in other countries are typically found to leave at higher rates than those whose parents have the experience of PSE, but who are often found to have lower leaving rates in Canada (Finnie & Qiu, 2008).

Second, predictive models can include many more factors or variables than might be used in a descriptive model, which is usually more parsimonious in order to tell a story of how the outcome of interest is related to a set of key explanatory variables. The predictive approach, thus, takes advantage of all the information available to maximize predictive accuracy.

Third, predictive models produce a specific probability value for each student for whom a prediction needs to be made. These probabilities will range across a continuous scale that runs from 0 to 1, where 0 means that the student will definitely not leave and 1 means he or she is certain to leave. This allows institutions to differentiate and order students at all risk levels — even though the probabilities generated by the model may ultimately be used to categorize students into different risk categories (e.g., high-, medium-, low-risk groups).

Predictive models and the individual-level student predictions they generate can, in the context of student retention, be used in a range of highly practical ways.[3] First, even though their first purpose is to provide predictions, they can also help an institution better understand how early leaving is related to the various student, program and other characteristics or factors included in the models.

Second, institutions can use the student-level leaving predictions generated by the models to target their student success initiatives (or other programs or activities) on those students who need them most.

Third, and related, the effects of any program targeted on students using such a set of predictions and related prediction cut-off points (e.g., those above a certain risk probability are targeted with the program while those below the cut-off are not) can be empirically estimated precisely due to the “discontinuity(s)” that characterizes such a targeting approach, and related estimation methods.[4]

Finally, the effectiveness of any student success program put in place in a more general way (e.g., one that is made available to all students) can be estimated across different student risk levels. For example, while it might make sense to target programs on students at the highest risk of leaving early, at least some programs might be most effective in improving outcomes for middle or even lower risk students, while other programs might work best for the most vulnerable. This kind of analysis can, therefore, help institutions develop a suite of programs that are most effective at improving student outcomes for the resources dedicated to them.

[3] To obtain the predictive model in the work reported here, we train our data using 2005 to 2012 cohorts by estimating several logistic regression models with different sets of predictors and do an external validation on 2013–14 cohorts. This is explained in more detail in Section 3.3.

[4] For more information on the effectiveness of regression discontinuity see Mayhew et al. (2016).

1.2 The Current Project: Using Predictive Models to Test Student Interventions

The project discussed in this report is based on using predictive models to first identify students' risk levels, and then to test three different student advising programs across three different student risk levels (low, medium, high) based on the predictions generated by the model.

This research is, therefore, placed not only in the broader context of the use of data analytics and predictive modelling, but also in the more specific context of the student retention literature, where student retention (the inverse of leaving a program before graduating) is generally understood to be the result of complex sets of factors, and difficult to measure (Wiggers & Arnold, 2011) and predict. At the same time, there is also a consensus among practitioners and researchers that there is no single solution to improving student success (Kuh, Kinzie, Schuh, & Whitt, 2005; Reason, 2009).

The concepts of predictive modelling — early-alert and intervention — and academic advising to support student success, therefore, form the foundations of this research project. This is the first of two reports looking at the relationship between predictive modelling, academic advising and student retention at Mohawk College.

This specific project is grounded in Mohawk College's Student Success Plan (2013), which identifies a system of support for students at risk of leaving college early. This system includes early intervention, comprehensive intrusive (proactive) advising, and student engagement through co-curricular and extra-curricular activities. Mohawk College's approach to support rests upon a foundation of post-admission (pre-arrival) assessments, predictive modelling and early intervention facilitated through advising.

The purposes of this phase of the research include first updating and evaluating the predictive model previously developed for Mohawk College by the Education Policy Research Initiative (EPRI) at the University of Ottawa using those students who entered Mohawk College from 2005 through 2012. The model is used to predict how likely each incoming student from the Fall 2013 and Fall 2014 cohorts was to leave the college early. These predictions are then used to test if the predictive model does an effective job of identifying student risk levels.

The predictive model is used to classify students into one of three equally sized Student Risk Classifications, or SRCs (low-risk, medium-risk, high-risk). It is further tested by comparing actual leaving rates across the three student risk categories.

The three risk groups are analyzed to identify the different student characteristics that tend to be represented in the student risk groups, while remembering that it is the full set of student characteristics that determine individual students' predicted leaving rates and, therefore, which risk group they will be in.

Finally, patterns of participation in current advising programs for students in the different risk groups are also calculated.

This modelling and these predictions and risk group assignments will then be used in the second phase of the research, which will test the efficacy of outreach and advising programs intended to improve student retention that have been put in place at Mohawk College. The results of the second research phase will be reported later this year.

2. Literature Review

2.1 Practices for Improving Student Outcomes

Many practices are cited in the literature as improving student success. For example, in the America College Testing's (ACT) *What Works in Student Retention* national survey, 96 items are listed as potential retention practices for practitioners to consider, implement and evaluate (Habley, Bloom & Robbins, 2012).

Kuh (2008) identified 10 high impact practices that significantly improve student learning in university environments, while the Center for Community College Student Engagement's (CCCSE) recent national report (2014) named 14 high impact practices related to improved student outcomes in the college environment. These are just a few of dozens of examples.

Furthermore, Tinto (1975, 1993), Braxton et al. (2004), Terenzini and Reason (2005), Reason (2009), and Braxton et al. (2014) have all contributed to a progressively evolving framework for student success. Braxton et al.'s (2014) theoretical framework recognizes the role of student entry characteristics, initial institutional commitment, external environment, internal environment and organizational characteristics, academic and intellectual development, and subsequent institutional commitment. Each of these elements has influence over the outcomes of student persistence. Such a broad collection of theories and practices demonstrate the complexity of student success.

However, within this literature — especially within the community college sector — a range of complementary activities are consistently identified as being promising for improving student outcomes. These include the use of predictive models (van Barneveld et al., 2012), early-alert and intervention practices (Center for Community College Student Engagement, 2014), and academic advising (Braxton et al., 2014).

Early alert and intervention programs are noted as high impact practices by the Center for Community College Student Engagement (2014), as these programs are considered to have the greatest effect on student retention. CCCSE defines early alert and intervention programs as a systematic process whereby instructors alert someone at the college when students in their classes are struggling academically, and that person contacts the students in an effort to get them the assistance they need.

This approach is consistent with the theory and practice of intrusive or proactive academic advising intervention (Glennen, 1975; Varney, 2013), which are purposeful approaches to outreach and support services. Proactive interventions are in stark contrast to the laissez-faire approach to student support services, which Dietsche (2012) found to be no longer relevant for supporting today's students. After an extensive survey of 60,000 Ontario college students, Dietsche (2012) concluded that proactive outreach and advising were critical to ensuring student success. The same argument was echoed by Poirier (2015) in his analysis of orientation and transition programs in three large Ontario colleges.

Citing examples of Habley et al. (2012), Braxton et al. (2014), and the Center for Community College Student Engagement (2014), Fricker (2015) recently argued that academic advising is commonly cited as a central service to foster student success, especially for community college students. The most recent example is a study by Braxton et al. (2014), which found empirical evidence to support a theory of student success on commuter campuses, but more noteworthy, academic advising was highlighted as one of the most important interventions. Since most Ontario colleges are commuter campuses, this theory and the recommendations for practice may have direct relevance.

Unfortunately, there is very little literature about student success, retention or the practice of academic advising on college campuses in Canada (Fricker, 2015). Academic advising is often defined in a broad way. Grites (1979, p.1) defined academic advising as “a decision-making process during which students realize their maximum educational potential through communication and information exchanges with an adviser.” Braxton et al. (2014) cited this definition in their recent work. Similarly, the Ontario Academic Advising Professionals (n.d.) stated that academic advising was “to be understood in its broadest sense and may include those involved in providing academic advice, career advice, counselling, liaison services, and/or learning skills opportunities in order to promote student success and retention.” These definitions are consistent with the role of the Student Success Advisers at Mohawk College and provide the context of the current research project.

This project presents a contemporary look at the relationship between participation in academic advising and student retention in college.

2.2 Data, Data Analytics and Predictive Modelling

There are a plethora of terms imprecisely related to practices that are connected to predicting student success in higher education. In an attempt to provide a common language for analytics in higher education, van Barneveld et al. (2012) provided a useful framework, which defined business analytics, academic analytics, learning analytics, predictive analytics and action analytics.

In this framework, “analytics is an overarching concept described as data-driven decision making” (van Barneveld et al., 2012, p. 6) with the aid of specialized computer systems, while business and academic analytics “allows management/executives access to indicators — historical or real-time through “dashboards” — of how the business (higher education institution) and its units (colleges, schools or departments) are performing.”

Building on these concepts, “predictive analytics is a process that serves all levels of higher education and business, and acts as a connector between the data collected, intelligent action that can be taken as a result of the analysis, and, ultimately, informed decision making.” They also proposed, more formally, that predictive analytics be defined as “an area of statistical analysis that deals with extracting information using various technologies to uncover relationships and patterns within large volumes of data that can be used to predict behaviour and events" (p. 8). Core to this concept is using analytics to take action and implement programs, services and interventions that support student success. Examples of projects listed by van Barneveld et al. include student success plans and student readiness inventories.

Various tools are available to universities and colleges to support this kind of work. A particularly promising approach involves using predictive models that can use past data to predict future outcomes of students at the individual level. A popular method is to use the logistic regression analysis of historical data to model retention patterns, which can then be used to predict the success of future cohorts. In this approach, one part of the student administrative data available (could be certain years of data or just a random selection of the whole sample) is used to develop the model, while the second part of the data (could be other years — typically more recent ones — or a random selection of the whole sample) is used to test the model and assess its performance.

The model with the lowest prediction error is then used to predict the individual probabilities of leaving college early for incoming students. As the predictive model takes into account the various student characteristics for which data are available, the model is more accurate in its prediction than predicting early leavers based on descriptive statistics alone (for example, by looking only at traditional risk indicators such as high school grades or performance on student assessment tests — or combinations of such indicators) and will, therefore, do a better job of predicting student success.

These kinds of predictive modelling approaches are used in many different fields in order to predict future outcomes based on historical patterns. The use of such models is perhaps most prevalent in finance, where predictive modelling is used to assess the risk of bankruptcy for individuals (Foster & Stine, 2004), and businesses (Atiya, 2001) based on their prior financial history, or in order to identify different types of fraud (Phua et al., 2010).

Similarly, predictive modelling is used in medicine to predict undiagnosed diseases or the prognosis of a diagnosed disease on the basis of patient traits (see, for example, Baan et al., 1999 and Federico et al., 2000). Predictive models are also a mainstay in certain engineering fields. For example, such models have been used to predict traffic accidents on busy highways (Hossain & Muromachi, 2012) or to implement power savings in vehicles based on predictive algorithms of driver behaviour (Murphey et al., 2008). Such algorithms are also used daily by various websites to provide targeted advertisements to their users based on individual web browsing habits (Perlich et al., 2014).

Research in education has been comparatively slow at adopting such approaches. To name a few, predictive models have been used to forecast grades (Kotsiantis, 2012), identify students at the risk of not finishing school on time (Aguiar et al., 2015, Lakkaraju et al., 2015, Sara et al., 2015), and student retention (Dekker et al., 2009, Delen, 2010, Lin, 2012, Nandeshwar et al., 2011, Thammasiri et al., 2014, Yu et al., 2010, Zhang et al., 2010).

However, to the best of the authors' knowledge, there are not many studies that test or validate a predictive model of student retention in Canada, aside from Conrad and Morris (2010) who analyzed student administrative data using the “random survival forest” machine-learning technique to predict student retention at York University.

Jia and Maloney (2015) empirically estimated the determinants of course non-completion outcomes in first year and student non-retention outcomes in second year using administrative data from a large public university in New Zealand.

3. Data and Methodology

3.1 Variable Definitions

This section briefly describes the variables used in the analysis discussed in this report. The selection of variables was determined by the availability of data at Mohawk College and follows a well-known and broadly used theoretical model in the persistence literature introduced by Tinto (1975, 1993). According to this model, students enter PSE with various pre-entry characteristics such as age, race, gender, family structure, parental educational attainment, and high school preparation, as well as their own skills and abilities. These factors contribute to the formation of students’ initial goals and their level of commitment to their studies. High school average data, which is explained below, is also included in the model used in this report.

Student and Program Variables

The set of student and program variables first includes the year of entry and gender. Age is also included and is broken down into six categories: below 18, 18, 19, 20–22, 23–26, and 27 and above. Regional status identifies whether a student resides in an urban or rural neighbourhood at the time of his or her application or whether the student is an international student (and thus does not have the preceding categories identified).

High school grade average tends to be one of the strongest predictors of retention for incoming students (Astin, 1997). This variable is computed as the average of the six highest grades from English and mathematics courses taken during the third and fourth year of high school. The categories for this variable are: A plus, A, A minus, B plus, B, C plus, C, D plus, D and F.

School corresponds to the program in which the student is enrolled (there are 17 schools). The credential variable has four categories: certificate (1-year), diploma (2-year), advanced diploma (3-year), and graduate certificate (1-year).

Student Entrance Survey Variables

Mohawk College administers a Student Entrance Survey (SES) for incoming students at the beginning of each term. This survey was introduced at Mohawk in 2006 as part of a ministry-funded project to administer the Ontario College Student Engagement Survey (Dietsche, 2009). The survey was developed by Peter Dietsche over the previous 20 years, including an earlier version, which was used in the Pan-Canadian Survey of College Students (Dietsche, 2007, 2008). The survey has been continuously used by Mohawk since 2006 and has gone through some editing and refinement.

The SES is not a mandatory survey, but it does include responses from around 70% of the incoming population of the Fall 2013 and Fall 2014 cohorts. It includes questions that are intended to help identify some of the risk factors for leaving college early. Student responses to specific questions or groups of questions define each of the following variables: low career clarity, low confidence in abilities, working 15 hours or more a week while studying, having a hard time transitioning to college life at Mohawk College and low educational commitment. These are all binary variables (At-Risk = 1 and Not-at-Risk = 0) that identify what are thought to be risk factors for leaving college early.

Assessment Scores

Incoming students at Mohawk College undertake writing, reading and mathematics assessments before the start of their first semester. The College Math Project (Orpwood, Schollen, Leek, Marinelli-Henriques & Assiri, 2012) and the College Student Achievement Project (2015) have studied and reported on the importance of these variables for student success in Ontario colleges over the last decade.

At Mohawk College, the reading and writing assessments are written on the Accuplacer platform and use the WritePlacer software to write and score the essay. The math assessment, developed by Mohawk College, is written on the Maple TA platform. Since there have been changes over the years in the scales on which these assessments are scored, each assessment variable was rescored to reflect the student’s relative position in the overall score distribution for the particular assessment he or she took.

The reading and mathematics assessment scores are grouped into eight categories which range from 1 to 8. The lowest category indicates that the student’s score is in the lowest end of the score distribution, while the highest category indicates that the student’s score is in the top end of the score distribution.

There are two categories, 1 and 2, for the writing assessment, representing below and above the median score.

In addition, since not all students take the assessment tests, missing categories were also included for each assessment. The proportion of students who completed the mathematics assessment is 34% and the proportion of students who completed the reading and writing assessments is 56%.[5]

[5] Not all incoming students are asked to take the math assessment test.

The Outcome Variables: Leaving Early and Participation in an Advising Program

The key outcome variable of interest in this analysis is whether or not the student left the program early. The leaving measure is binary (did not leave = 0 and left = 1) and represents retention from the first to the second term (first-term retention).

Day 10 of each term is used as the date to identify student retention, which corresponds to the end of the "add/drop" period at Mohawk College. In other words, this is the last day in the term that students can register. Students who are registered on day 10 of the initial fall term are included in the analysis, and students are considered to have remained at Mohawk College if they are again registered at day 10 of the second term in the winter (first-term retention) or day 10 of the third term (one-year retention) in the following fall semester.[6]

Participation in student advising is the second outcome of interest. The relevant variable is defined as whether the student sought the help of Student Success Advisers at least once during the semester. These data are collected through an advising software program called ClockWork, which is utilized by all Student Success Advisers at Mohawk College. For the purposes of this analysis, this is also a binary variable (Did not Seek Advising = 0, Sought Advising = 1). This variable does not account for the frequency of advising support provided to a student, or the length or type of advising interaction that occurred.

3.2 Sample Populations and Restrictions

The data used to estimate the retention model underlying the student leaving predictions included students who entered Mohawk College from 2005 to 2012.

Predicted leaving rates were then generated for Mohawk College students who entered the college in Fall 2013 and Fall 2014.

All analysis is restricted to students at the Fennell (main) campus, because this is where the advising programs of interest in this project were put in place in Fall 2015.

3.3 The Predictive Model

The predictive model used in this analysis was initially developed in previous work, but was updated for the current project. A logistic regression model approach is employed, which is used frequently in the broader economics literature to model binary (0–1) outcomes such as going to college, being unemployed and migration decisions (to name but a few examples).

[6] The withdrawal status of those who return for the second semester is checked and those who withdraw before day 10 are also considered leavers. Some students who leave (particularly those who do not withdraw) will not have an entry in this file, but simply do not show up the next semester. Our approach allows for students who leave Mohawk College but do not go through the formal withdrawal process to be identified as leavers.

In the logistic regression model in our case, the probability of leaving early is defined as:

$$\boldsymbol{P} = \frac{\boldsymbol{e^{\beta_0+\beta_1 X}}}{\boldsymbol{1 + e^{\beta_0+\beta_1 X}}}$$

The terms $\boldsymbol{\beta_0}$ and $\boldsymbol{\beta_1}$ represent the parameters associated with each element of $\boldsymbol{X}$, the set of predictor variables included in the model (listed above), and denote the individual effect on leaving of each variable taking account of the other variables included in the model.[7]

The final choice of the predictive model was based on a comparison across many different model specifications, which included different combinations of background, risk and assessment variables. For example, one specification included only the student and program variables (i.e., gender, urban-rural status, age, credential, high school average and school). This model was extended to include interaction terms among the different sets of variables. These models were then also expanded to include the risk and assessment variables both separately and together.

Data for the Fall 2005 to 2012 cohorts were used as the training data for developing our predictive model. We use the Fall 2013–14 cohort data to do an external validation. The performance of each model specification was compared based on the log loss value it generated. Log loss essentially quantifies the extent of how much the predicted probabilities of leaving (values between 0 and 1, based on Fall 2005 to 2012 data) differ from the actual leaving outcomes (binary based on Fall 2013–14 data).[8] The model specification with the lowest log loss value was chosen as the best predictive model.

3.4 Predicted Leaving Rates

The coefficient estimates from this model were then used to generate the predicted probability of leaving early for each student from the Fall 2013 and 2014 cohorts. Given the coefficient estimates of the predictive model (i.e., the $\hat{\beta}$'s), each student's individual and program characteristics, SES-based risk factors, and assessment scores were entered into the model-based formula above to obtain the individual prediction for the probability of leaving before graduating.

The mathematical formula implied by the logistic regression model for obtaining individual predictions for probability of leaving college early ($\hat{P}$) is:

$$\frac{e^{\hat{\beta}_0+\hat{\beta}_1\times Gender+\hat{\beta}_2\times Age+\hat{\beta}_3\times Cred+\hat{\beta}_4\times Sch+\hat{\beta}_5\times Urb+\hat{\beta}_6\times HSavg+\hat{\beta}_7\times Assess+\hat{\beta}_8\times Risk}}{1+e^{\hat{\beta}_0+\hat{\beta}_1\times Gender+\hat{\beta}_2\times Age+\hat{\beta}_3\times Cred+\hat{\beta}_4\times Sch+\hat{\beta}_5\times Urb+\hat{\beta}_6\times HSavg+\hat{\beta}_7\times Assess+\hat{\beta}_8\times Risk}}$$

[7] The term *e* represents the base of the natural logarithm, which corresponds approximately to 2.718.

[8] The formula for log loss is $-\frac{1}{N}\sum_{i=1}^{N} p_i \log(\hat{p}_i) + (1-p_i)\log(1-\hat{p}_i)$, where the terms p_i and $\hat{p_i}$ denote the actual leaving outcome and the predicted probability of leaving for student i, respectively.

The individual predicted probability for each student is given on a scale of 0 to 1.0 (or 0 to 100%). These predicted probabilities are (as discussed in the introduction) centred around the overall leaving rate across the different cohorts.

Most importantly, students with individual and program characteristics, SES scores and assessment scores identified by the model to be associated with a greater likelihood of leaving tend to have higher predicted probabilities of leaving, and vice versa. But it is the entire set of characteristics that determine each student's own predicted probability of leaving, which is the essential attribute and strength of a modelling approach.

It must also be recognized that some students with high predicted probabilities do not, in fact, leave, and the reverse is equally true. This is inherent in the predictive nature of what is essentially a statistical exercise.

4. Predictive Model Performance

4.1 Estimation Results for the Predictive Model

The results of the logit model used to generate the predicted leaving rates are shown in Appendix Table A.1. We will not go into these results in any detail here because the precise details of the underlying model are not central to the current paper, which is instead focused on the predictions the model generates.

That said, predictions are only as good as the models upon which they are based, and so a brief discussion of the model estimation results is warranted. The table presents the logit parameter estimates. Due to the non-linear nature of the logit model, these have no straight-forward intuitive meaning, but the direction of the effects and statistical significance of the parameter estimates are meaningful.

The main findings are as follows:

- Females are significantly less likely to leave than males
- Older students (i.e., those aged 23–26 and especially those 27 or older) leave at lower rates than younger students
- Region based on the student's address at the time of application to Mohawk (urban, rural, international, missing) does not matter
- Certificate students have higher leaving rates, and graduate certificate students leave at lower rates than advanced diploma students (the omitted/comparison group)
- Students with lower incoming grades, especially those with a D plus or lower, leave at higher rates
- There are substantial differences in leaving rates across schools
- Of the SES-based risk categories, "career clarity" and especially "educational commitment" are significant predictors of which students leave early
- Reading and math assessment scores matter

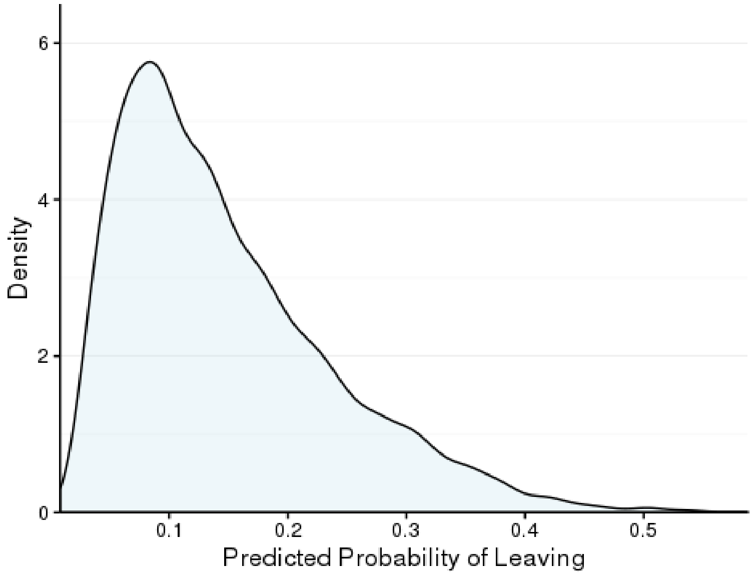
4.2 The Predictions: Model Performance

The Predicted Leaving Rates

The predictive model was then used to generate the predicted probability of leaving for each student in the 2013 and 2014 cohorts, as described above. The resulting predicted leaving rates are shown in Figure 1, which represents the probability density function of the predictions. On the horizontal axis of Figure 1 is the range of possible predicted leaving probabilities, which runs from 0 to 1.0. The vertical axis essentially represents the proportion of students as they are distributed across those different levels, and adds up to one (i.e., thus capturing the entire sample).

In practice, the observations range from very low predicted leaving probabilities to an effective maximum of about .5, with most being under .3, which shows that most students have fairly low predicted probabilities of leaving. This is not surprising in a context where just 15.8% of those in the estimation sample (2005 to 2012 cohorts) and 15.3% of students in the test sample (2013–14 cohorts) actually leave. The peak of the distribution is around .10, and the distribution is skewed right.

Figure 1: Probability Density of Leaving Early



Risk Levels and Actual Leaving Rates

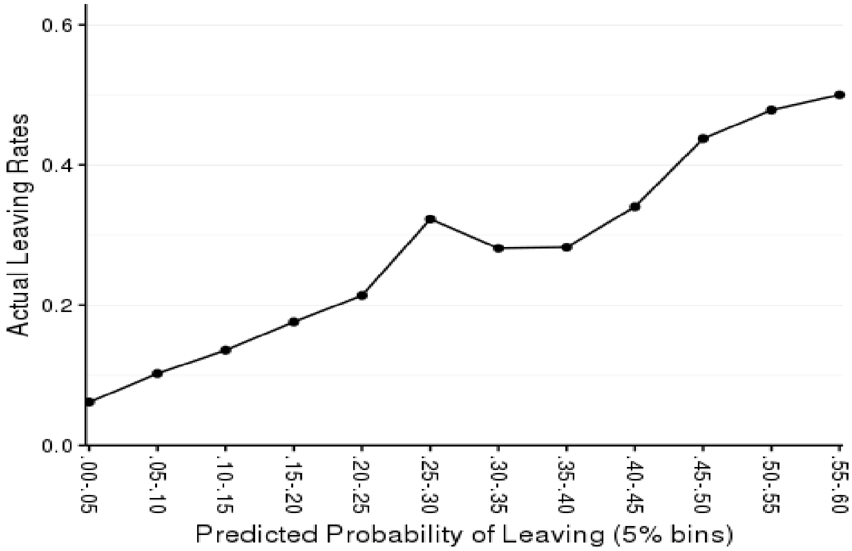
One way of assessing a model's ability to accurately predict leaving rates is to compare the individual predicted probabilities of leaving generated by the predictive model to actual leaving rates. To do so, the 2013 and 2014 cohorts for which leaving rates were predicted, were divided into 20 groups defined by their predicted risk levels, and these were compared to the actual leaving rate for each of these groups (which is, of course, known).

Figure 2 shows that the predicted risk values are strongly correlated with the actual leaving rates. Each value on the x-axis represents the group of students according to their risk values as given by the predictive model, and the corresponding y-values represent the actual leaving rates for these groups. That is, students assigned a low-risk value by the model actually leave at lower rates, and those who are assigned a high-risk value leave at higher rates.

This type of strong positive relationship would be expected for in-sample predictions (i.e., predictions computed on the estimating sample, the 2005 to 2012 data in our case), as the model is estimated to fit the actual leaving rates. However, such a strong positive relationship is not always guaranteed for out-of-sample predictions (i.e., predictions computed on a new sample, the 2013–14 data in the present case). Figure 2: Distribution of Predicted Probabilities of Leaving versus Actual Leaving Rates thus shows that the predictive model does a very good job of ordering students based on their tendencies to leave college early.

These results also suggest that the model should do well in ordering new incoming students according to their risk levels, so that student support programs can, for example, be effectively targeted, which is one of the principal purposes of developing such a predictive model.

Figure 2: Distribution of Predicted Probabilities of Leaving versus Actual Leaving Rates



Predicting Leavers and Stayers

Another way to assess the performance of a predictive model is to check the percentage of total correct predictions (also called accuracy), false positive and false negative predictions. The computation of these statistics depends, however, on choosing a threshold value that assigns each student a binary predicted leaving outcome (i.e., they are predicted either to leave or stay).

For example, if the threshold is set at 50%, then students whose individual predicted probabilities of leaving are greater than 0.50 are predicted to be leavers and those whose probabilities are 0.50 or lower are predicted to be stayers. The resulting predicted leaving outcomes are then tabulated against actual leaving outcomes (still for the 2013 and 2014 test cohorts for which we both generate predictions and ultimately know if the student left or stayed).

There is, however, no single correct threshold value to use for categorizing students in this way. Furthermore, any cut-off will arbitrarily assign individuals around the cut-off who have almost identical predicted probabilities of leaving (say, .29 vs. .31 if .30 is selected as the cut-off) into one category or the other (i.e., leaver or stayer), when really there is only a slight difference in the probability of their leaving. There is, in short, nothing special about any single cut-off or the predictions it generates, and any chosen threshold has problems.

While a cut-off of 50% may seem intuitively appealing to some, this will not, in general, be a good threshold for dividing individuals into those who are predicted to leave and those who are predicted to stay. Considering there is a limited set of predictors available and that the resulting predictions tend to cluster to the left, reflecting the relatively low overall leaving rates being explained by the model, a 50% cut-off would identify very few students as leavers.

In general, when choosing a cut-off (or threshold) there is a trade-off between the total correct predictions, and false positive and negative predictions. Choosing a higher probability-of-leaving threshold (or cut-off) to identify predicted leavers and predicted stayers will generally result in (as mentioned above) underestimating the overall number of leavers (since fewer students will have predicted probabilities above that threshold) and will, in particular, categorize many actual leavers as predicted stayers. The reverse will hold if the threshold is set too low.

It is, therefore, customary to use different thresholds to test a model and answer the question: “How good is the model at accurately predicting who will stay and who will leave?” Appendix Table A.2 shows i) total correct predictions, ii) false positive predictions and iii) false negative predictions for different threshold values. Correct predictions are obvious: those who are predicted to leave do so and the same holds for stayers. The false positive predictions represent the cases where students are predicted to be leavers but actually stayed, and false negative predictions represent the cases where students are predicted to be continuers, but actually left.

For the chosen model, setting a threshold level of 10% generates many false negative predictions and a relatively small number of false positive predictions. As Table A.2 depicts, for a higher threshold value such as 20% or 30%, there are many more false negatives, but fewer false positive predictions — as fewer students overall are predicted to leave at the higher thresholds.

The threshold to be used in any such exercise will ultimately depend on how the resulting predictions will be used. That is, choosing a preferred cut-off becomes a policy decision as much as a statistical exercise and will depend on the estimated costs of erring on one side or the other: false positives versus false negatives, or over-predicting how many will leave versus how many will stay.

Both imply costs for the institution: a false positive is a false alarm and potentially leads to a needless and costly intervention for students who don’t really need help. A false negative, on the other hand, represents an at-risk student who flies under the radar and who does not receive help when he or she should, and perhaps drops out as a result. The ideal cut-off balances these costs and ultimately depends on many factors, including the effectiveness and costs of available interventions.

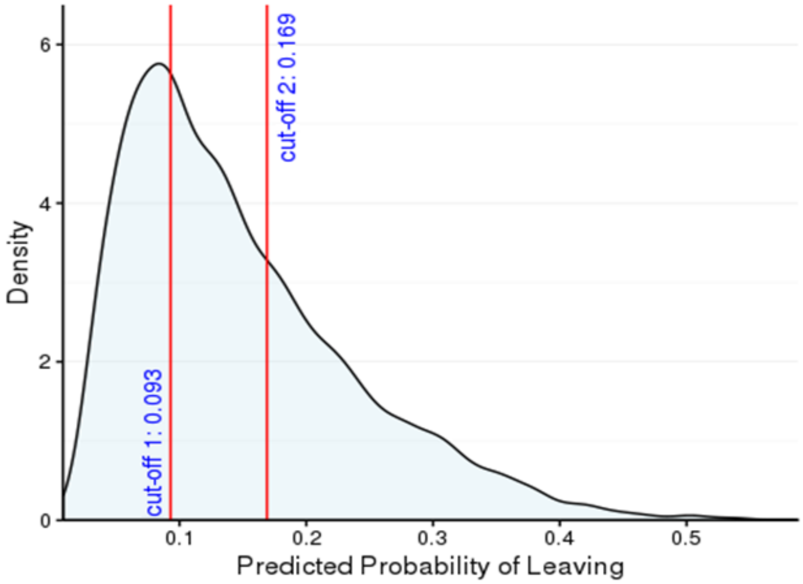
5. Student Risk Classifications (SRCs)

One of the objectives of the broader research project is to test the different student advising strategies put into place for the incoming 2015 cohort of students across different student risk levels. The full distribution of risk levels (or predicted leaving rates) was therefore divided into three ranges using two cut-off points to do so (Cut-off1<Cut-off2).

Figure 3 again shows the distribution of predicted probabilities of leaving college early for the 2013 and 2014 entry cohorts together. The cut-off points were chosen so that the students were equally divided across three Student Risk Classifications (SRCs); that is, each group constitutes 33.3% of the student population.

Students whose predicted probabilities of leaving are less than or equal to Cut-off1 are assigned to the low-risk group; students whose predicted probabilities are higher than Cut-off1 and lower than Cut-off2 are assigned to the medium-risk group; and those whose predicted probabilities are higher than or equal to Cut-off2 are assigned to the high-risk group.

Figure 3: Distribution of Predicted Probabilities of Leaving



Since the distribution of predicted probabilities of leaving is skewed to the right, and the maximum value for the individual predicted probability of leaving as generated by the predictive model is 58.7%, the cut-offs that give this equal distribution of students are not very high. Cut-off1, which identifies the low-risk group, is 0.093 (or a predicted probability of leaving of 9.3%) and Cut-off2, which separates the medium and high-risk groups, is 0.169 (or 16.9%).

Table 1 presents the leaving rates for the three SRCs for each of the test cohorts of students (2013 and 2014). The table shows increasingly higher leaving rates going from the low-risk group to high-risk group, as would be expected.

Taking the two cohorts together, the proportion of leavers in the high-risk group (24%) is substantially higher than the overall leaving rate (16%) and, of course, higher than the medium-risk group (14%) and, especially, the low-risk group (just 9%). In this way, the table also provides another perspective of the accuracy of the predictive model and its effectiveness in ordering students by risk level.

Table 1: Leaving Rates (%) and Number of Leavers (N) by Student Risk Classification

	2013		2014		All	
Risk Level	Rate	N	Rate	N	Rate	N
Low	9	163	8	145	9	308
Medium	13	250	15	246	14	496
High	23	431	26	412	24	843
Total	15	844	16	803	16	1647

Note: Cut-off 1=9.3%, Cut-off 2=16.9%

5.1 Analysis of the SRCs: Which Students are in Which Risk Groups?

What These Results Represent

Table 2 through to Table 15 present the distribution of students across the three SRCs for the 2013 and 2014 cohorts combined (the results are very similar for the two cohorts) across the range of variables representing student and program characteristics, SES risk indicators and assessment scores.

While many of the variables considered include certain categories that individually tend to be associated with leaving early (being male, having low grades, etc.), the predictive model sorts out the relationships between leaving early and each individual factor while taking all other factors into account. As a result, some factors that may appear to be important when viewed in isolation (as below) may not be significant when included in the more general model (while the reverse can also occur).[9]

Generating the individual student predicted risk levels then involves taking into account all the information on students and assigns them an overall risk value based on the predictive model.

Finally, the SRCs then order students by risk level into the low-risk, medium-risk and high-risk categories as described above.

Student and Program Variables

Table 2 shows that the gender distribution differs across SRCs. Females represent 47% of the overall student population for the 2013 and 2014 cohorts together, while the proportion of females is only around 36% for

[9] This is simply due to correlations between the predictors. Suppose we have a simple regression model that includes only gender and find that men leave at higher rates than women. If we add field of study, this gender gap will be reduced if men tend to be in fields of study with higher leaving rates and the initial gender effect is due at least in part to the correlation of gender with this other risk factor.

the high-risk group and is 56% for the low-risk group. In other words, females tend to exhibit a lower tendency to leave early.

Table 2: Gender Distributions (%) by Student Risk Classification

	Low	Medium	High	All
Male	44	50	64	53
Female	56	50	36	47
Total	100	100	100	100

Table 3 shows that compared to the medium- and high-risk groups, the low-risk group has a higher proportion of older students. Around 34% of students who are age 23 and above are in the low-risk group, while around 21% are in the medium-risk group and 15% are in the high-risk group. That is, older students tend to have a lower tendency to leave college early.

Table 3: Age Distributions (%) by Student Risk Classification

	Low	Medium	High	All
0-17	5	6	4	5
18	19	26	21	22
19	13	22	27	21
20-22	28	25	33	29
23-26	19	11	10	13
27 and above	15	9	5	10
Total	100	100	100	100

As evidenced in Table 4, the distribution of students in terms of international, urban and rural status is not considerably different across SRCs. The medium-risk group has slightly lower proportions of international students and higher proportions of urban students compared to the other two groups. For the 2013 and 2014 cohorts together, around 3% of the medium-risk group are international students, while around 6% of the low- and high-risk groups are international students. The low-risk group has a lower proportion (82%) of domestic students who live in urban areas compared to the medium- and high-risk groups which have 86% and 85%, respectively.

Table 4: Regional Status Distributions (%) by Student Risk Classification

	Low	Medium	High	All
Rural	11	11	10	10
Urban	82	86	85	84
International	6	3	6	5
Total	100	100	100	100

As shown in Table 5, there is a clear pattern in how high school average grade distributions differ by SRC. Looking at the 2013 and 2014 cohorts together, around 13% of the low-risk students attained high school averages of A minus and above, while only 4% of high-risk students maintained those averages. Conversely, around 29% of high-risk students had high school averages of D plus and below, while only 6% of low-risk students had an average below C. The medium-risk group has a high school average distribution that is very similar to the overall distribution, with the majority of students concentrated between C and B plus.

Table 5: High School Average Distributions (%) by Student Risk Classification

	Low	Medium	High	All
F	1	2	6	3
D	1	3	10	4
D plus	4	9	13	9
C	9	15	15	13
C plus	14	19	14	16
B	17	17	12	15
B plus	14	12	8	11
A minus	8	6	3	6
A	4	2	1	2
A plus	1	1	0	1
Missing	26	15	18	19
Total	100	100	100	100

Table 6 shows the sample distributions by school. Around 31% of low-risk students are enrolled in programs in the School of Justice and Wellness, while around 2% of high-risk students are enrolled in these programs. Around 9% of the low-risk students are enrolled in graduate certificate programs in the School of Business, whereas there are no students in that school within the medium- and high-risk groups. In addition, around 22% of the high-risk students are enrolled in the School of Interdisciplinary Studies and 27% are enrolled in

Management Studies, while only 1% and 8% of the low-risk group are enrolled in programs within these schools.

Table 6: School Distributions (%) by Student Risk Classification

	Low	Medium	High	All
Building and Construction Sciences	7	6	5	6
Chemical and Environmental	4	5	5	5
Computer Science and IT	7	5	3	5
Electrotechnology	2	3	6	4
Engineering Technology Preparatory	0	1	4	2
Graduate Studies (Business)	9	0	0	3
Human Services	9	19	10	13
Interdisciplinary Studies	1	5	22	9
Justice and Wellness Studies	31	14	2	16
Management Studies	8	13	27	16
Mechanical and Industrial Technology	4	4	6	5
Media and Entertainment	10	16	7	11
Office Administration Studies	6	6	1	4
Skilled Trades	2	2	2	2
Total	100	100	100	100

As Table 7 reveals, the distribution of students in terms of credentials also differs across SRCs. Within the high-risk group, around 33% of students are certificate students while there are no high-risk graduate certificate students. Within the low-risk group, around 17% of students are graduate certificate students and certificate students constitute only around 3% of students. In other words, graduate certificate students are predicted to be at a much lower risk of leaving than those obtaining a regular certificate.

Table 7: Credential Distributions (%) by Student Risk Classification

	Low	Medium	High	All
Certificate	3	12	33	16
Diploma	52	58	43	51
Advanced Diploma	28	30	23	27
Graduate Certificate	17	0	0	6
Total	100	100	100	100

Student Entrance Survey (SES) Variables

As mentioned above, the risk variables based on Mohawk College's Student Entrance Survey (SES) questions represent hypothesized individual indicators of a student's risk of leaving college early based on those variables considered independently and one at a time. In contrast, the predictive model takes into account all the information on students, including not only the SES risk variables, but also all the other factors included in the model and, therefore, best indicates which variables, including the SES risk indicators, are the best predictors of leaving early.

Looking at the SES risk markers — "career clarity," "confidence," "educational commitment" and "transition" — in Table 8 to 11, of those students who did take the SES, those who are classified as medium- or high-risk students on the SRC are also more likely to be labelled at risk by the relevant SES variable.

Table 8: At Risk (Career Clarity) Distributions (%) by Student Risk Classification

	Low	Medium	High	All
No	69	68	37	58
Yes	8	11	10	10
Missing	23	21	52	32
Total	100	100	100	100

Table 9: At Risk (Confidence) Distributions (%) by Student Risk Classification

	Low	Medium	High	All
No	59	53	29	47
Yes	18	26	19	21
Missing	23	21	52	32
Total	100	100	100	100

Table 10: At Risk (Educational Commitment) Distributions (%) by Student Risk Classification

	Low	**Medium**	**High**	**All**
No	77	77	44	66
Yes	1	2	4	2
Missing	23	21	52	32
Total	100	100	100	100

Table 11: At Risk (Transition) Distributions (%) by Student Risk Classification

	Low	Medium	High	All
No	66	66	38	57
Yes	11	13	10	11
Missing	23	21	52	32
Total	100	100	100	100

In contrast to other SES variables, among the students who took the SES, the proportion of students working 15 hours or more (which is taken as an indication of being at-risk of leaving college early) does not differ significantly across the SRCs (Table 12).

Table 12: At Risk (Working 15+ Hours) Distributions (%) by Student Risk Classification

	Low	**Medium**	**High**	**All**
No	43	46	28	39
Yes	34	33	20	29
Missing	23	21	52	32
Total	100	100	100	100

Assessment Scores

There are also differences in the distribution of the mathematics, reading and writing assessment scores across different SRCs. The high-risk group tends to have higher proportions of students who scored poorly compared to other groups, while the low-risk group tends to have higher proportions of students who performed well on these tests. In addition, the high-risk group has significantly higher proportions of students without reading and writing assessment scores. The following section presents results in more detail on how the distributions differ across SRCs.

As shown in Table 13, the low-risk group has a higher proportion (around 28%) of students who scored within the top three categories (6, 7 or 8) compared to the high-risk group (less than 1%). In addition, the high-risk group has a higher proportion (around 24%) of students who scored in the bottom three categories (1, 2 or 3) compared to the low-risk group (less than 2%).

Table 13: Math Assessment Score Distributions (%) by Student Risk Classification

	Low	Medium	High	All
1	0	2	10	4
2	0	4	8	4
3	1	5	6	4
4	2	6	4	4
5	4	6	2	4
6	8	4	0	4
7	9	3	0	4
8	11	1	0	4
Missing	63	68	69	67
Total	100	100	100	100

Table 14 depicts the reading assessment score distributions by SRCs. Around 36% of students in each of the medium- and low-risk groups do not have reading scores, while 60% of the high-risk students do not have assessment scores. The low-risk group has a higher proportion of students who scored in the top three categories compared to the high-risk group. The high- and medium-risk groups have a higher proportion of students (around 26%) who scored in the bottom three categories compared to the low-risk group (around 11%).

Table 14: Reading Assessment Score Distributions (%) by Student Risk Classification

	Low	Medium	High	All
1	2	8	11	7
2	4	9	8	7
3	5	9	7	7
4	7	9	5	7
5	8	9	4	7
6	10	8	2	7
7	13	7	2	7
8	14	5	1	7
Missing	36	36	60	44
Total	100	100	100	100

Table 15 presents the writing assessment score distributions by SRC. Similar to reading scores, around 36% of students in the medium- and low-risk groups do not have writing scores, as opposed to around 60% for high-risk students. For the high-risk group, the majority (27%) of the remaining 40% of students scored in the bottom category. For the low-risk group, around 23% of the remaining students, whose score is non-missing, scored in the bottom category.

Table 15: Writing Assessment Score Distributions (%) by Student Risk Classification

	Low	Medium	High	All
1	23	33	27	28
2	41	31	13	28
Missing	36	36	60	44
Total	100	100	100	100

5.2 Advising Participation Rates

Figure 4 shows the general relationship between risk levels and participating in student advising programs. The advising is binary and simply identifies who sought assistance from an adviser and who did not. Therefore, this analysis does not account for the frequency of advising support provided to a student, or the length or type of advising interaction that occurred. However, the hope is that students in need of advising programs (i.e., at-risk students) are the ones who actually participate in them. The results seem to show just that; a general correlation between the predicted probability of leaving and advising participation rates. In other words, the higher the risk level, the greater the proportion of students participating in advising programs. There are a few potential reasons for this relationship. The first is that high-risk students thought they needed the advising services the most and took advantage of the opportunities presented.[10]

Figure 4: Predicted Probabilities of Leaving versus Advising Participation Rates

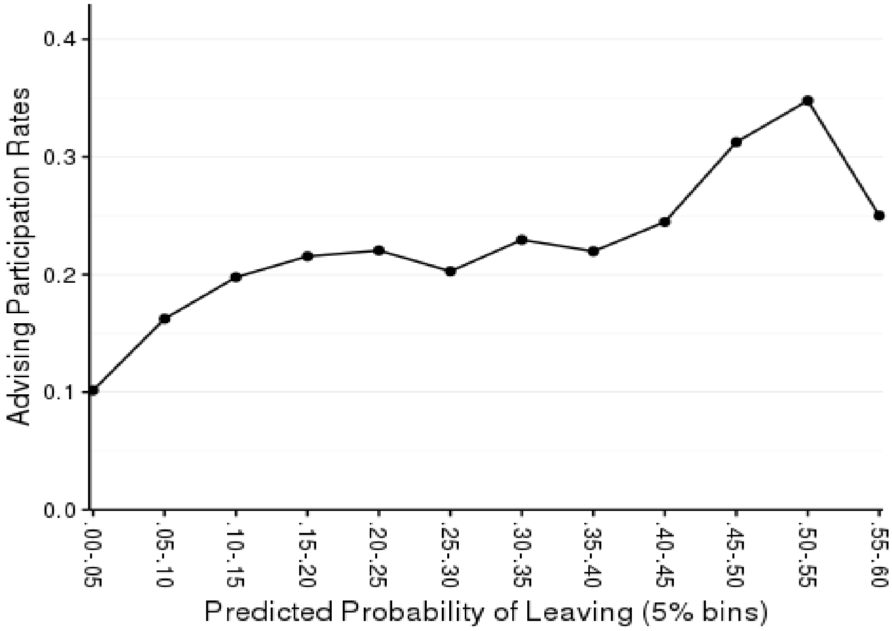


Table 16 depicts students that seek advising across SRCs. Medium- and high-risk groups have the highest advising participation rates, while the low-risk group has the lowest rates.

[10] In a survey of over 60,000 college students in Canada, Dietsche (2012) found that those students who say they need support services often do not actually access them. This kind of evidence is a driving factor behind the theory and practice of intrusive (or proactive) advising. This is a common approach employed by Mohawk College Student Success Advisers, and this active outreach to students deemed at-risk of leaving may be the reason for the relationship observed in the data. This will be explored further in Phase 2 of this project.

Table 16: Advising Participation Rates and Sample Sizes by Student Risk Classification*

	2013		**2014**		**All**	
Risk Level	Rate	N	Rate	N	Rate	N
Low	16	279	12	208	14	487
Medium	21	390	20	323	21	713
High	22	424	21	324	22	748
Total	20	1093	17	855	19	1948

Note: The advising participation rate is calculated as the ratio of the number of students who sought academic advising services at least once to the total number of students in the semester.
**Cut-off 1=9.3%, Cut-off 2=16.9%*

The overall advising participation rate is around 3 percentage points lower for the Fall 2014 cohort (17%) when compared to the Fall 2013 cohort (20%). The overall participation rate for 2013 and 2014 combined is around 20%.

The difference in advising participation rates for the medium- and high-risk groups is only around a percentage point, while the low-risk group has a participation rate that is around 5 to 7 percentage points lower than the medium-risk group and around 6 to 9 percentage points lower than the high-risk group.

6. Discussion

Developing and Using a Predictive Model to Identify Students at Risk of Leaving Early

This paper represents the first phase of a broader research project that seeks to estimate the effects of three different approaches to student advising randomly assigned to all incoming students before their entry to Mohawk College in Fall 2015. The differences in these approaches include varying degrees of intrusiveness, or how the students are contacted as well as the nature of advising services offered, including one-on-one versus group sessions.

One of the purposes of the broader project is to estimate the effects of the different advising initiatives not only overall, but also across students at different risk levels in terms of their likelihood of leaving Mohawk College early (without graduating). These assessments will then allow Mohawk College to better understand and target these initiatives in the future and ultimately obtain the greatest possible improvements in student outcomes relative to the costs of these different initiatives.

To identify students' risk levels of leaving early, a predictive modelling approach has been adopted. This work builds on previous collaborative projects focused on student retention undertaken by Mohawk College and the Education Policy Research Initiative (EPRI).

A logit model specification is employed, whereby the probability that any given student will leave early is related to a range of factors, including a student's demographic and program characteristics (gender, age, region, high school grades, school and credential); a set of risk indicators developed by Mohawk College from its Student Entrance Survey (SES) relating to things such as career clarity, emotional readiness, and the number of hours worked in outside jobs; and reading, writing and math assessments also carried out before students enter Mohawk College.

For the current project, the models that were developed in the earlier work have been updated, re-specified and tested further. Those tests have included, first, comparing a wide range of model specifications using the estimation sample (i.e., the 2005 to 2012 entry cohorts of Mohawk College students) and using standard econometric techniques (i.e., log loss functions) in order to arrive at the model that best fits the data.

Further tests have been carried out to assess the accuracy of the predictive model in predicting leaving rates using a separate test data set (i.e., the 2013 and 2014 entry cohorts). In one set of tests, the predicted probabilities of leaving (which range from 0 to 1.0 for any given student) were compared to actual leaving rates to see how well the predicted probabilities track actual leaving rates. Other tests involved turning students' predicted leaving rates into individual-level predictions of who will persist and who will leave and again comparing these to who actually stays and who leaves.

The model performed well on both counts, even though it is based only on variables readily available in Mohawk College's administrative records, as listed above.

The individual-level predicted leaving rates were then used to assign students — again those in the 2013–14 test cohorts — into three risk categories. These risk categories were, for convenience, arbitrarily chosen to be of equal size (i.e., each category comprised a third of the incoming students). These were first used to further test the predictive power of the predictive model by comparing actual leaving rates across the low-, medium- and high-risk categories to actual leaving rates, and again these tracked well.

Once students were assigned to one of the risk groups, a descriptive analysis was undertaken to identify the distribution of students across these groups according to the various student and program characteristics, SES risk indicators and assessment scores mentioned above.

Of course, students with the individual risk factors identified by the predictive model (e.g., being male, having low high-school grades, possessing any of the SES-based risk factors defined by Mohawk College, having low assessment scores) were found in the higher risk groups identified by the predictive model. But many were not.

Specific Findings

Some of the specific findings are as follows:

- The main determinants of leaving Mohawk College before graduation as identified by the predictive model are as follows: females, students aged 23 or older, and graduate certificate students have lower leaving rates than others; certificate students and those with lower incoming grades (especially D plus or lower) leave at higher rates; there are substantial differences in leaving rates across schools; region (urban, rural, international) is not statistically significant; of the SES-based risk categories, only “career clarity” and especially “education commitment” are significant; reading and math assessment scores also matter.

- The individual-level predicted student leaving rates (with leavers defined as students who did not start the second semester of their program) generated by the predictive model, which was estimated over the 2005 to 2012 entry cohorts and then tested for the 2013 and 2014 entry cohorts, closely tracked actual leaving rates.
- The actual leaving rates of students in the three student risk categories defined using the predictive model are 24% for high-risk students, 14% for medium-risk students, and 9% for the low-risk students in the 2013–14 test cohorts, showing how the predictive model effectively differentiates students according to their risk levels.
- Finally, contrary to common perceptions of student affairs practitioners that the students who participate in advising and support programs are those who do not actually need them (Dietsche, 2012), it was found that the high- and medium-risk groups participated the most in existing student advising programs at about 22%, while only 14% of the low-risk students participated (all results again for the 2013–14 test cohorts).

All of this work sets up Phase 2 of the research project, where the alternative types of advising interventions put in place for the 2015 entry cohort will be evaluated and their effects tested across different student risk levels.

Implications and Lessons Learned for Mohawk College

One general finding of importance for Mohawk College is that the predictive model approach creates a continuum of predicted leaving rates for students, with the majority of students grouped in a relatively narrow band of predicted probabilities centred on the actual overall leaving rates for the different cohorts — around 16% for the first-term leaving rate, which is the focus of this paper.

As a result, changing the cut-off applied to the full distribution of predicted probabilities used to divide students into risk groups will result in different sized groups with different characteristics. This has implications for operational planning such as identifying the number and characteristics of which students are to be targeted with any given student success initiative. Other cut-offs could be used and other groups could be identified, with there being nothing sacred (or indeed very special) about the three groups identified here.

The research project also caused Mohawk College to abandon its initial terminology of “College Ready,” “Underprepared” and “At-Risk” students and instead adopt the use of high-, medium-, and low-risk students.[11] Further work and consideration are needed to determine the best terms to describe the different categories to be used, whether these are three in number, four, or some other number, and whatever predicted leaving cut-offs are used to define them.

One measure that could be put in place is more intentional and intense outreach to students in order to improve student outcomes, which would be consistent with the college's focus on proactive advising.

Another potential measure, focused on better measurement of students' risk levels, could be to require participation in the Student Entrance Survey through various institutional policy levers, such as embedding it in the acceptance or registration process.

Testing, evaluating and reporting on any of these new approaches could generate meaningful evidence on effective outreach practices and policies. Similarly, the additional data provided by students would provide a more complete picture of student needs, strengthen the predictive model and inform early intervention support services.

Additionally, the SES could probably be more useful if it used scales to represent the relevant risk factors rather than the binary approach currently used, since a scale would provide more information. In fact, it makes little sense to first create sets of binary "student risk indicators" based on the more detailed underlying information available and then include those indicators in the predictive models. A better strategy would be to enter the raw SES (scaled) variables directly into the models. Furthermore, the risk indicators should be empirically verified rather than based on *a priori* expectations of what may indicate that a student is at risk.

An additional implication of the findings is that the college needs to examine entrance requirements to ensure that they have the desired effect. In particular, having a high school average of D plus or below represents an important set of variables in the predictive models. Such students were more likely to be in the high-risk group once all their characteristics were taken into account. Therefore, minimum entrance requirements and/or conditional entrance should be studied to determine if such policy changes could lead to improved retention rates by informing the design of appropriate interventions or programs to support students with low high-school grades. Additional access programs may also be better suited for students within this category.

Further research is also needed to better understand students' intentions and academic goals. While early leaving rates are substantial, some students may from the beginning not intend to complete their studies. For example, some students may be using entry into a Mohawk College program as a stepping stone to another postsecondary education experience. Better understanding student goals would, therefore,

[11] While the initial description assisted the college in establishing an understanding of the basic concept, high, medium and low are better descriptors. Most notably, College Ready may incorrectly imply that there is no risk and At-Risk may imply that there is no risk associated with the other classifications.

improve Mohawk College's planning and could provide useful information concerning program upgrading, laddering and pathways to other PSE institutions.

Finally, gaining an understanding of why students leave college early (and what their plans are) would complete the picture of student success and retention. In both of these circumstances, leaving the college may be deemed a success by the student, which is a story that should also be told.

Limitations of the Model and Further Opportunities

The analysis presented here has some general limitations worth mentioning. First, in our case, the results of the predictive model were based on how students behaved over the period of time covered in the estimation (2005 to 2012), and then tested over the 2013 and 2014 cohorts (other approaches for creating, training and testing samples can be used). If student behaviour has changed since that time (e.g., individuals drop out at generally different rates or at relatively different rates across groups or specific student characteristics in ways that are not captured by the model), the model will no longer reflect current behaviour. This is inevitable for all predictive models that are trained on past data and tested on future data.

Second, the predictive model provides the means of targeting students for student success initiatives at the individual level and evaluating the effectiveness of programs put in place. While these statistical exercises can inform future policies in extremely important ways (allowing decisions to be evidence based), they do not and cannot determine the institution's actual policies themselves. In other words, an institution's policies will ultimately be driven by its overall or specific objectives, the resources available and other potential factors.

The third limitation of the analysis undertaken here is that the performance of any predictive model can only be as good as the data provided. This analysis has, in particular, been limited to using the information currently being obtained by Mohawk College before students get to college. The very good reason for this is that Mohawk College wants to target its students early on, before the start of the first semester.

A first step in terms of data collection would be to make sure all relevant data currently available at Mohawk College are being used in the predictive model. Student financial aid (loan) data or program choice data from college applications might be examples of data that are at least potentially available and which could be used. The Student Entrance Survey information used in the model could also be included in its original, more detailed format rather than the binary form represented by the risk factors created by Mohawk.

A second step might be to use the available data to create additional variables. For example, a proxy of socio-economic background could be created by using postal code information provided by students to link to other data sources (such as the census) to identify the characteristics of the neighbourhood where the student was living before he or she entered the college (e.g., average income, education levels, housing, language).

Third, if advising services (or other student success strategies) were offered after the semester started (or in a later semester), early student performance could also be included in the models, which would undoubtedly improve their performance substantially. Examples of these include early course reporting (even just attendance would likely be useful), interim grades and final grades for predicting retention in subsequent semesters.

Finally, more and more electronic information is available on students and could be included in any analysis of student retention and then used to predict student retention and help target and measure interventions. A good example is student online course participation. Most courses now have an important online component, or at least require some online participation on the part of the student, and this information could be mined to both better understand student outcomes and to better predict which students are at highest risk of leaving or are at imminent threat of doing so.

Phase 2 of the Project

The second phase of this project seeks to evaluate a new outreach and advising initiative at Mohawk College designed to provide greater support to all new students attending the Fennell campus in the fall of 2015. All students randomly received one of three early outreach and advising packages prior to the start of classes.

Avenues for Further Research

Predictive models of student retention are useful for three main purposes:

- To help an institution better understand how early leaving is related to the various student, program and other characteristics/factors included in the models
- To use the individual, student-level leaving predictions generated by the model to target student success initiatives or other student-focused programs or activities
- To estimate the effects of student success (or other) initiatives targeted using a predictive model approach by the use of discontinuity analysis, which is an estimation approach based precisely on the sorts of cut-offs that can be used to target students, or to estimate the effects of other initiatives across different student risk levels

The first point is, as indicated, more oriented toward better understanding student retention, and purely “descriptive” models, which can and generally should be different from “predictive models,” will generally be the preferred approach if that is the sole objective. But the general approach is broadly the same: using statistical models to relate leaving (or retention) to various factors of interest. Descriptive models are differentiated by their being specified in a way that tells a story or (as the term implies) describes student retention — and stop there.

Predictive models are, on the other hand, less concerned with telling a story, and are instead focused on developing specifications that best fit the data and generate the most accurate predictions of student outcomes. They may, for example, be a bit messier than purely descriptive models, such as including more variable interactions. This can make for not as clean a picture of student retention — but can generate better predictions.

We suggest that more modelling of both types is key to advancing our understanding of student retention as it currently is, and to targeting and testing new initiatives aimed at improving student outcomes.

We would thus hope to see, first of all, more PSE institutions engaged in these modelling exercises as the foundation for better understanding their students and student outcomes, and developing better student-focused policies using an evidence-based approach.

Secondly, current models can almost always be improved, especially by adding information/data to the models and we have sketched out a few directions for this in the preceding sub-section, including:

1. Adding additional variables based on information already available at institutions, including those related to financial aid, or program choice data from college applications, as well as utilizing some of the information already included in the models more thoroughly, such as the data contained in the Student Entrance Survey
2. Creating additional variables by linking to other data sources (e.g., the census to capture socio-economic background) from the existing data (i.e., postal code)
3. Adding "early reporting" on students to better understand and predict student behaviour and outcomes after their point of entry into the institution
4. Diving into the potentially massive sources of electronic information increasingly being collected on students, such as those related to course participation and engagement

Thirdly, predictive models come in a variety of forms. The one employed in this project is based on a fairly simple logit model approach, but other approaches can be tried and compared as in some of the work in education research, mainly in the US, cited in the literature review section of this report. We see advanced machine-learning algorithms as representing a particularly important avenue for new work, although these approaches need to be more thoroughly developed and tested before their efficacy is determined. They also have the distinct disadvantage of being much more complex than simpler modelling approaches of the type used here, which can represent a cost at a number of different levels — including their use on a wider basis to institutional practitioners who might want to develop, update and apply predictive models on an on-going basis.

Finally, predictive models can be put to much greater use in targeting and testing student support initiatives as also mentioned in other studies cited in this text, such as Delen (2010) and Zhang et al. (2010).

We see great opportunities in all of these directions in the PSE context. It is hoped that this paper might contribute to that progress.

Definitions

Data analytics are processes of data assessment and analysis to inform decisions at all levels, for example institutions, organizations and businesses (van Barneveld et al., 2012).

High-risk category represents the students who constitute the top third of the predicted probability of leaving distribution. This group of students is least likely to succeed and more likely to leave college after one term and, therefore, needs the most support and intervention. At the beginning of the project, this group was named "At-risk."

Intrusive/proactive advising represents the intrusive approach to advising introduced by Glennen (1975), which calls for more deliberate advising interventions and supportive advising relationships to enhance student motivation. More recently, this approach has been known as proactive advising (Varney, 2013), which encourages proactive outreach, offering support before a student needs it, and fostering strong relationships between an adviser and a student.

Lifecycle advising corresponds to the different advising services offered at the many cyclical elements of a student's experience in college. From semester to semester or year to year, there are predictable activities, deadlines, challenges and experiences that occur routinely (i.e., registration, payment deadlines, midterms and finals) and that constitute the lifecycle of a college student. The various activities that recognize, respond to and support a student through those common experiences constitute lifecycle advising.

Low-risk category represents the students who constitute the bottom third of the predicted probability of leaving distribution. This group of students is most likely to succeed and least likely to leave college after one term and, therefore, needs the least support. At the beginning of the project, this group was named "College Ready."

Medium-risk category represents the students who constitute the middle third of the predicted probability of leaving distribution. This group of students is in the middle in terms of being likely to succeed and/or leave college after one term. At the beginning of the project, this group was named "Underprepared."

Predictive analytics, a type of data analytics, are a set of technologies used to uncover relationships and patterns within large volumes of data that can be used to predict behaviour and events (van Barneveld et al., 2012).

Predictive model is an output of the predictive modelling process that is used to predict an outcome of interest given the values for predictor variables. In the context of this project, it refers to the regression model specification used to predict the probability of a student leaving college after one term. These predictions were also used to determine the cut-off points for identifying the low-, medium- and high-risk categories (groups) of students.

Predictive modelling, encompassed in predictive analytics, is a set of mathematical techniques used to find a relationship between an outcome or dependent variable, and predictor or independent variables to predict the unknown or new values of the dependent variable (Dickey, 2012).

Student Entrance Survey (SES) is the post-admission, pre-registration survey the majority of new students take as part of their transition to Mohawk College. It occurs on campus at the same time as the Assessments for Success (AFS). AFS are the reading, writing and mathematics post-admission, pre-registration placement tests new students take during their transition to college. Based on the results of their assessments, students are placed in either regular or remedial communications or mathematics courses. They are also provided resources and support for any upgrading necessary prior to the start of classes.

Student Risk Classifications (SRCs) are the low-, medium- and high-risk categories of students across the full distribution of risk levels (or predicted leaving rates) identified in this research project. The cut-off points were chosen so that the students were equally divided across the distribution and each group had 33.3% of the student population.

References

Aguiar, E., Lakkaraju, H., Bhanpuri, N., Miller, D., Yuhas, B., & Addison, K. L. (2015, March). Who, when, and why: A machine learning approach to prioritizing students at risk of not graduating high school on time. In Proceedings of the Fifth International Conference on Learning Analytics and Knowledge (pp. 93-102). ACM.

Astin, A. W. (1997). How “good” is your institution’s retention rate? *Research in Higher Education*, *38*(6), 647–658.

Atiya, A. F. (2001). Bankruptcy prediction for credit risk using neural networks: A survey and new results. IEEE Transactions on neural networks, 12(4), 929-935.

Baan, C. A., Ruige, J. B., Stolk, R. P., Witteman, J. C., Dekker, J. M., Heine, R. J., & Feskens, E. J. (1999). Performance of a predictive model to identify undiagnosed diabetes in a health care setting. Diabetes Care, 22(2), 213-219.

Braxton, J. M., Doyle, W. R., Hartley III, H. V., Hirschy, A. S., Jones, W. A., & McClendon, M. K. (2014). *Rethinking college student retention*. San Francisco, CA: Jossey-Bass.

Braxton, J. M., Hirschy, A. S., & McClendon, M. K. (2004). Understanding and reducing college student departure. *ASHE-ERIC Higher Education Report*, *30*(3).

Center for Community College Student Engagement. (2014). *A matter of degrees: Practices to pathways (High-impact practices for community college student success).* Austin, TX: The University of Texas at Austin, Program in Higher Education Leadership.

College Student Achievement Project Team. (2015). *College student achievement project: Final report 2015*. Toronto, ON: Seneca College of Applied Arts and Technology. Retrieved from http://csap.senecacollege.ca/docs/CSAP%20Cycle%202%20final%20report%2011Jun15.pdf

Conrad, M., & Morris, K. (2010). *Shifting from retention rates to retention risk: An alternative approach for managing institutional student retention performance*. Toronto, ON: Higher Education Quality Council of Ontario.

Dekker, G., Pechenizkiy, M., & Vleeshouwers, J. (2009, July). Predicting students drop out: A case study. In Educational Data Mining.

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

Dickey, D. A. (2012). Introduction to Predictive Modeling with Examples, SAS Global Forum 2012, Paper 337.

Dietsche, P. (2007) Pan-Canadian Study of First Year College Students: Student Characteristics and the College Experience, Report 1. Gatineau, QC: Association of Canadian Community Colleges/Human Resources and Social Development Canada. Retrieved from https://pseinfosys.com/wp-content/uploads/2014/02/Dietsche-ACCC-2008.-Pan-Canadian-Study-of-First-Year-College-Students-1.pdf

Dietsche, P. (2008) Pan-Canadian Study of First Year College Students: The Characteristics and Experience of Aboriginal, Disabled, Immigrant and Visible Minority Students, Report 2. Gatineau, QC: Association of Canadian Community Colleges/Human Resources and Social Development Canada. Retrieved from https://pseinfosys.com/wp-content/uploads/2014/02/Dietsche-ACCC-2008.-Pan-Canadian-Study-of-First-Year-College-Students-2.pdf

Dietsche, P. (2009). The Ontario College Student Engagement Survey 2006-2009: Final report-project results, data modelling, tests of reliability and validity and future directions. *Toronto, ON: Prepared for the Ontario Ministry of Training, Colleges and Universities*.

Dietsche, P. H. J. (2012). Use of Campus Support Services by Ontario College Students. *Canadian Journal of Higher Education*, *42*(3), 65–92.

Federico, M., Vitolo, U., Zinzani, P. L., Chisesi, T., Clò, V., Bellesi, G., & Pavone, V. (2000). Prognosis of follicular lymphoma: a predictive model based on a retrospective analysis of 987 cases. Blood, 95(3), 783-789.

Finnie, R., & Qiu, T. (2008). The patterns of persistence in post-secondary education in Canada: Evidence from the YITS-B dataset. Educational Policy Institute.

Foster, D., & Stine, R. (2004). Variable Selection in Data Mining: Building a predictive model for bankruptcy. *Journal of the American Statistical Association*, 99(466), 303–313. Retrieved from https://doi.org/10.1198/016214504000000287

Fricker, T. (2015). The relationship between academic advising and student success in Canadian colleges: A review of the literature. *College Quarterly*, *18*(4).

Glennen, R. E. (1975). Intrusive college counseling. *College Student Journal*, *9*(1), 2–4.

Grites, T. J. (1979). *Academic Advising: Getting Us Through the Eighties* (AAHE-ERIC Higher Education Research Report No. 7). Washington, DC: American Association of Higher Education.

Habley, W. R., Bloom, J. L., & Robbins, S. (2012). *Increasing Persistence: Research-based strategies for college student success*. San Francisco, CA: Jossey-Bass.

Hossain, M., & Muromachi, Y. (2012). A Bayesian network based framework for real-time crash prediction on the basic freeway segments of urban expressways. Accident Analysis & Prevention, 45, 373-381.

Jia, P., & Maloney, T. (2015). Using predictive modelling to identify students at risk of poor university outcomes. Higher Education, 70(1), 127-149..

Kotsiantis, S. B. (2012). Use of machine learning techniques for educational proposes: A decision support system for forecasting students' grades. Artificial Intelligence Review, 37(4), 331-344.

Kuh, G. D. (2008). *High-impact educational practices: What they are, who has access to them, and why they matter*. Washington, DC: Association of American Colleges & Universities.

Kuh, G. D., Kinzie, J., Schuh, J. H., & Whitt, E. J. (2005). *Student success in college: Creating conditions that matter*. San Francisco, CA: Jossey-Bass. Retrieved from http://www.loc.gov/catdir/toc/ecip054/2004027912.html

Lakkaraju, H., Aguiar, E., Shan, C., Miller, D., Bhanpuri, N., Ghani, R., & Addison, K. L. (2015, August). A machine learning framework to identify students at risk of adverse academic outcomes. In Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 1909-1918). ACM.

Lin, S. H. (2012). Data mining for student retention management. *Journal of Computing Sciences in Colleges*, 27(4), 92-99.

Mayhew, M., Rockenbach, A., Bowman, N., Seifert, T., Wolniak, G., Pascarella, E., and Terenzini, P. (2016) *How College Affects Students: 21st Century Evidence that Higher Education Works, Volume 3*. San Francisco, CA: Jossey-Bass, pp. 523-574.

Mohawk College. (2013). *Mohawk's 5 Point Student Success Plan*. Hamilton, ON.

Murphey, Y. L., Chen, Z., Kiliaris, L., Park, J., Kuang, M., Masrur, A., & Phillips, A. (2008, June). Neural learning of driving environment predictions for vehicle power management. In 2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence) (pp. 3755-3761). IEEE.

Nandeshwar, A., Menzies, T., & Nelson, A. (2011). Learning patterns of university student retention. *Expert Systems with Applications*, 38(12), 14984-14996.

Ontario Academic Advising Professionals. (n.d.). Terms of Reference. Retrieved December 13, 2015, from http://oaap.ca/terms-of-reference/

Orpwood, G., Schollen, L., Leek, G., Marinelli-Henriques, P., & Assiri, H. (2012). *College mathematics project 2011: Final report*. Toronto, ON: Seneca College of Applied Arts and Technology. Retrieved from http://collegemathproject.senecac.on.ca/cmp/en/pdf/FinalReport/2011/CMP_2011_Final_Report%20-%2002Apr12%20pmh.pdf

Perlich, C., Dalessandro, B., Raeder, T., Stitelman, O., & Provost, F. (2014). Machine learning for targeted display advertising: Transfer learning in action. Machine learning, 95(1), 103-127.

Phua, C., Lee, V., Smith, K., & Gayler, R. (2010). A comprehensive survey of data mining-based fraud detection research. arXiv preprint arXiv:1009.6119.

Poirier, W. J. (2015, June). *Deliberate orientation and transition practices as part of a broader student success strategy: A comparative study of three large urban Ontario colleges* (thesis). Ontario Institute for Studies in Education, University of Toronto, Toronto, ON. Retrieved from http://hdl.handle.net/1807/69456

Reason, R. D. (2009). An Examination of Persistence Research Through the Lens of a Comprehensive Conceptual Framework. *Journal of College Student Development*, *50*(6), 659–682.

Sara, N. B., Halland, R., Igel, C., & Alstrup, S. (2015). High-school dropout prediction using machine learning: A Danish large-scale study. In Proceedings (p. 319). Presses universitaires de Louvain.

Terenzini, P. T., & Reason, R. D. (2005). Parsing the first year of college: Rethinking the effects of college on students. Presented at the Annual Conference of the Association for the Study of Higher Education, Philadelphia, PA.

Thammasiri, D., Delen, D., Meesad, P., & Kasap, N. (2014). A critical assessment of imbalanced class distribution problem: The case of predicting freshmen student attrition. Expert Systems with Applications, 41(2), 321-330.

Tinto, V. (1975). Dropout from higher education: A theoretical synthesis of recent research. *Review of Educational Research*, *45*, 89–125.

Tinto, V. (1993). *Leaving college: Rethinking the causes and cures of student attrition* (2nd ed.). Chicago, IL: University of Chicago Press.

van Barneveld, A., Arnold, K. E., & Campbell, J. P. (2012). *Anaytics in higher education: Establishing a common language* (Educause Learning Initiative No. 1) (p. 11). EDUCAUSE. Retrieved from https://library.educause.edu/resources/2012/1/analytics-in-higher-education-establishing-a-common-language

Varney, J. (2013). Proactive Advising. In J. K. Drake, P. Jordan, & M. A. Miller (Eds.), *Academic Advising Approaches: Strategies that teach students to make the most of college*. San Francisco, CA: Jossey-Bass.

Wiggers, R., & Arnold, C. (2011). *Defining, Measuring and Achieving "Student Success" in Ontario Colleges and Universities* (@ Issue Paper No. No. 10). Toronto, ON: Higher Education Quality Council of Ontario. Retrieved from http://www.heqco.ca/SiteCollectionDocuments/AtIssueStudent%20Success%20ENG.pdf

Yu, C. H., DiGangi, S., Jannasch-Pennell, A., & Kaprolet, C. (2010). A data mining approach for identifying predictors of student retention from sophomore to junior year. *Journal of Data Science*, 8(2), 307-325.

Zhang, Y., Oussena, S., Clark, T., & Kim, H. (2010). Using data mining to improve student retention in higher education: A case study. In International Conference on Enterprise Information Systems.

Higher Education
Quality Council
of Ontario

An agency of the Government of Ontario