Vaughan van Appel
Department of Statistics, University of Johannesburg, South Africa
Email: vvanappel@uj.ac.za

Rina Durandt*
Department of Mathematics and Applied Mathematics, University of Johannesburg, South Africa
Email: rdurandt@uj.ac.za

DOI: http://dx.doi.org/10.18820/2519593X/pie.v37i2.1
ISSN 0258-2236
e-ISSN 2519-593X
Perspectives in Education
2019 37(2): 1-15

Date Published:
27 November 2019

INVESTIGATING POSSIBILITIES OF PREDICTIVE MATHEMATICAL MODELS TO IDENTIFY AT-RISK STUDENTS IN THE SOUTH AFRICAN HIGHER EDUCATION CONTEXT

ABSTRACT

This article reports on the results of an investigation of the predictive accuracy of five different mathematical models to identify at-risk students in a Business Statistics course. Low levels of students' success, especially in mathematics-related subjects such as statistics, are a salient problem in South Africa and other countries. Statistical knowledge is included in a variety of programmes offered by many faculties at tertiary level, and early prediction of at-risk students seems necessary to enhance academic success especially when dealing with large class groups. In this study, we used 395 Business Statistics students' grades from an academic semester at an urban university in South Africa to build a predictive model to identify at-risk students. Grounded on Meyer's model evaluation criteria and striving for a balance between accuracy and simplicity, two out of five models are identified as viable predictive models in identifying at-risk students by using a cross-validation test. The article shows the possibilities and limits in deriving information from a number of covariates. These results are interesting and have implications for educational practice in statistics courses.

Keywords: *At-risk Students; Large Classes; Predictive Modelling; Mathematical Model; Statistics Courses; Tertiary Education.*

1. INTRODUCTION

The transition from school to university seems exciting for many first year students, but a daunting fact is the high drop-out rates, especially in mathematics-related courses. This phenomenon is locally and globally a matter of concern and universities are making efforts to investigate the reasons and eliminate the occurrence of this phenomenon (compare Greefrath, Koepf & Neugebauer, 2016; Van Zyl, Gravett & De Bruin, 2012). Research results published by The Centre for Development and Enterprise in South Africa (SA) confirmed a significant underperformance in education at school level, particularly in mathematics teaching and learning (Bernstein, 2013). Rach and Heinze (2017) reported

that the transition from school to university, particularly in mathematics, is often a substantial hurdle for students. Results from their research, conducted in Germany on 182 first-semester university students majoring in mathematics, indicated only a marginal influence of school-related mathematical resources on first semester study success. They explain mathematical related courses at university is very different from school mathematics and the learning cultures at both institutions could be a reason for the transition problems between school and university. At tertiary level, a clear distinction exists between statistics education and mathematics education, although both subjects fall under the umbrella mathematical sciences.

In the current SA context, statistics forms some part of the mathematics school curriculum. Bernstein's report (2013) underlined three key factors, among others, that are noteworthy for the teaching and learning of mathematics at school level (that also includes the teaching and learning of statistics at school level), namely (i) poor mathematics teachers' competencies (related to content and pedagogy); (ii) poor mathematics students' competencies; and (iii) a large gap in mathematics competencies among school students from the lowest income areas (approximately 66% of the population) and those from the richest areas. To appreciate the scale of mathematics schooling deficiencies in SA and the challenges that lie ahead, Bernstein's report (2013) puts learners' and teachers' competencies into an international context and further highlight significant disparity within SA (between quintile[1] 5 schools from the richest areas and quintile 1, 2, and 3 schools reflecting the average rural area, small towns, and most townships). For example, the competency of SA Grade 6 mathematics teachers are placed at the bottom end of the spectrum compared to a selection of other Eastern and Southern African countries. Mathematics is a key requirement for not only entry into higher education, but also for most modern, knowledge-intensive work (Bernstein, 2013). The report underlines that pass rates at universities are low, with an eventual graduate rate of roughly half the students at contact education universities who start a bachelor's degree. One concludes from these results a large number of SA students, who enter formal tertiary education, might be underprepared in terms of fundamental mathematical content for mathematics-related courses, such as statistics.

Successful transition between school and university is dependent on factors such as knowledge related to scientific mathematics and students' abilities to develop adequate learning strategies (Rach & Heinze, 2017). It seems logical that the latter factors are also important for learning statistics at tertiary level. Garfield and Ben-Zvi (2007: 380-381) argue that learning statistics involves an integration of a first "statistical literacy" component, a second "statistical reasoning", and a third "statistical thinking" component. The first component, statistical literacy, is often the expected outcome of introductory courses in statistics and is generally described as an understanding of the basic terms, symbols, tools of statistics, and the recognising and interpretation of different representations of data. Garfield and Ben-Zvi (2007) emphasised that statistics students' understanding of the basic concepts of statistics can easily be underestimated or overestimated by the educator. The second component, statistical reasoning, refers to the way people reason with statistical ideas and make sense of statistical information. The third component, statistical thinking, involves a higher order of thinking than statistical reasoning, more the way professionals will think. Ideally, all three components should be integrated to increase proficiency and for educators to

1 In South Africa the term 'quintile' is used as the national poverty ranking of public schools and their learners. Five different groups exist into which all public ordinary schools and their learners are placed – from the poorest in quintile 1 to the least poor in quintile 5.

assess student's understanding of statistics. In this inquiry, students were expected to master the statistical literacy component and show some evidence of statistical reasoning, but they were not expected to act as professionals. We argue that every student can learn statistical literacy and develop basic reasoning skills to ultimately reach academic success – the focus should be to meet the students' needs and not the incapacity of students.

Almost three decades ago, an *at-risk* student was defined as "one who is in danger of failing to complete his or her education with an adequate level of skills" (Slavin & Madden, 1989:4). Educators often view at-risk students as the ones who are more likely to fail than pass the course, and the term may be applied to students who face particular circumstances (e.g., low test scores, or low class attendance) that could jeopardise their ability to be academically successful. A broad overview of the literature revealed numerous studies on students' lack of academic success or delay during formal tertiary education, particularly in numeracy related fields, with a number of contributing factors (Cassidy, 2015; Greefrath et al., 2017; Onwuegbuzie, 2004; Rach & Heinze, 2017). Some of these contributing factors include a lack of self-efficacy, statistical anxiety, fostering negative attitudes towards statistics courses, and students finding statistics content difficult (Coetzee & van der Merwe, 2010; Onwuegbuzie, 2004; Talsma, Schüz, Schwarzer, & Norris, 2018; van Appel & Durandt, 2018). In addition, Science and Engineering courses generally obtain lower pass rates than many other courses at tertiary level, making it very important to take part in the global discussion regarding academic success, the efforts to identify at-risk students as early as possible and the concerning variables or factors to support these students on their academic path. Within the SA context, but also globally, educators (lecturers, faculties, and universities) are continuously more strained to increase pass rates and at the same time present students with a quality course. It therefore seems essential to identify the needs of students as early as possible to improve instructional practice. However, the focus of this inquiry is merely to identify at-risk students in a statistics course at a public university in SA, and not to improve instructional practices as such, although the latter might be seen as a possible outcome of this investigation.

The two research questions are: (1) what is a suitable predictive mathematical model used to identify at-risk students in a Business Statistics course at tertiary level, and (2) how effective is such a model to predict students' academic success in this course? In answering these research questions, we attempt to broaden our knowledge about the identification of at-risk students in a statistics course as early on as possible in the academic semester, and to identify and improve a suitable mathematical model that can provide trustworthy results in an educational context. These results have implications for the educational practice; it could contribute towards promptly detecting the needs of at-risk students, and ultimately resulting in enriched instructional practices and improved throughput rates in statistics.

2. PRIMARY THEORETICAL PERSPECTIVES

2.1 Theoretical guidelines for learning statistics

Learning statistics is grounded in the learning of mathematics, although it has developed as a research area in its own right with a growing network of researchers studying the development of students' statistical literacy, reasoning, and thinking. According to Garfield and Ben-Zvi (2007) research studies focus not only on statistics instruction, but also on the development of conceptual understanding. Kilpatrick, Swafford and Findell (2001) defined

five different strands of mathematical knowledge, which in combination indicate mathematical proficiency. These strands (supported by Samuelsson, 2010: 62) seem to connect particularly well with learning statistical knowledge (literacy, reasoning, and thinking) and are therefore relevant to this study, and include:

I. *Conceptual understanding* – the ability to grasp mathematical and/or statistical concepts, operations, and relationships.

II. *Procedural fluency* – the skill of performing flexible procedures accurately, efficiently, and appropriately to support mathematics and/or statistics learning.

III. *Strategic competence* – the ability to formulate, represent, and solve mathematical and/or statistical problems to contribute in developing various mathematical and/or statistical competencies and appropriate attitudes.

IV. *Adaptive reasoning* – the capacity for logical thought, reflection, explanation, and justification, in order to contribute to an adequate picture of mathematics and/or statistics.

V. *Productive disposition* – the ability to view mathematics and/or statistics as sensible, useful, and worthwhile, together with a belief in self-confidence.

Both Boaler (2000) and Samuelsson (2010) argued that situational context is a key aspect in producing mathematical knowledge and seems important to learn statistical literacy, which forms the foundation for reasoning and thinking. Garfield and Ben-Zvi (2007) claim statistical literacy is often seen as an expected outcome of schooling and a necessary component of adults' numeracy and literacy. In summary, we argue that the five strands of mathematical knowledge, by Kilpatrick et al. (2001), provide the knowledge base for students to learn statistics. Apart from the knowledge base, a notion of supporting disposition is also present with the belief that every student can attain the necessary skills for academic success when formal education meets the students' needs and not the incapacity of students.

2.2 Mathematical models and their relevance for teaching

Doerr, Ärlebäck and Misfeldt (2017: 71) underscored the substantial impact of mathematical models at all levels of society by claiming that "mathematical models are used to control processes, to design products, to monitor and influence economic systems, to enhance human agency, and to structure and understand the natural world in society and above all in the workplace". Earlier, Lesh and Doerr (2003) described models as conceptual systems that are used for some specific purpose. It seems reasonable to consider a mathematical model to monitor study success in an educational context, but if such a model is used to improve decision-making (e.g. regarding study success) then the quality of the model is also important. In the context of this study, we used a mathematical model in the context of teaching to identify students that are at-risk of failing a Business Statistics course. The idea was to implement the model as early as possible in the second academic semester of the academic year. Thus, the emphasis was more on the product and its efficiency as it would play a role in decision-making in the education context, than on all the steps taken during a modelling cycle. Meyer (2012: 150 – 222), proposed six criteria for mathematical models that should be considered during the evaluation of a model: (i) accuracy, (ii) realism, (iii) precision, (iv) robustness, (v) generalisability, and (vi) fruitfulness.

We followed the traditional steps in a modelling cycle; to identify the problem in the real world (e.g. low throughput rates in mathematical related courses), to make assumptions and identify variables by selecting relevant information and finding relations (e.g. using

continuous and formal assessment marks), to formulate a mathematical model and perform procedures to find results (e.g. multiple regression, or linear regression, or decision trees), to analyse and assess the solution by questioning the results and consequences (e.g. by checking the model's accuracy), and to iterate the modelling process to refine and extend the model (e.g. to use different data sets) (compare Blum and Leiβ, 2007; COMAP-SIAM, 2016). Through this modelling process we purposefully considered the evaluation criteria from Meyer (2012): (i) *accuracy*, if the output values were correct or near correct; (ii) *realistic*, if the model is based on correct assumptions; (iii) *precise*, if its predictions were in definite numbers, and imprecise, if its predictions were in a range of numbers; (iv) *robust*, if the model is to some extent protected against errors in the input data; (v) *general*, if it applied to a variety of educational contexts; and (vi) *fruitful*, if it resulted in useful conclusions. In this study we used a predictive model for a first-year Business Statistics course, but such a model can easily be adapted and used in other courses, given that there is enough reliable data available to 'train' the model. By predictive modelling we intended to use mathematical and computational methods to predict students' probability of academic success based on changes in the model input values. A genuine evaluation of the accuracy of the model used in this study will require observations over a number of years, or perhaps observations in a variety of courses. A limitation in building accurate predictive models is that these models are largely dependent on the quality of the data available. Thus, in this study, the quality and reliability of the data (which were only collected in one academic semester) will influence the accuracy (or predictive power) of the model. We attempted to increase the reliability and consistency of the data through a well-structured course plan with multiple assessment opportunities. Data from future research could refine the mathematical models and its implications in other educational context, for example to investigate the 'optimal' time to implement an at-risk model in an academic semester.

3. RESEARCH DESIGN

Model building seems important in identifying at-risk students, and in this study the accuracy of models was studied to determine the most suitable predictive model to identify at-risk students in a Business Statistics course offered at the University of Johannesburg during the second semester of 2017. Data were collected from 395 first year (undergraduate) students. In particular, we used this data to build five predictive models to predict the possible outcome of a future students' success in the course. Moreover, the data used in the building of the predictive models are often referred to as the training data. The aim of training a predictive model is for the model to learn to identify patterns or trends from the training data, which is then used in predicting the success of future students in the course. The ever-growing need to increase throughput rates and maintain high course standards makes predictive modelling especially useful. For example, in large classes, a predictive model can be used to identify students that are likely to fail the course and provides the educator with information about students that require much needed assistance. This identification process would be tedious in a large class without a predictive model, and would only be possible later in the academic semester.

For the training data, we used the 2017 gradebook that consisted of each students online cumulative average quiz mark and, two formal semester test marks, which accumulated to a final period mark (FPM). Thereafter, the students wrote an examination (EM), which was weighted in equal parts to compute the final mark (FM). In order for a student to pass the course, the student must obtain a FPM and EM greater or equal to 40%, and then a FM

greater or equal to 50%. In addition to the gradebook, we also collected the grade the students obtained in their prerequisite module, if students were repeating the module (Yes or No), the gender of the students, and the student's high school quintile ranking (range between quintile 1 and 5). More specifically, in our model building design we considered the school quintile ranking to be a descriptive of the student's socio-economic status. Socio-economic status generally combines three measures based on income, education, and occupation. Therefore, the school quintile raking represents a good predictor for the socio-economic variable in this study, as a high quintile ranking (levels 4 and 5) indicates a well-performing school (normally these schools are situated in or around major cities, where tuition fees are paid) indicating a higher socio-economic status. In contrast, low quintile ranked schools (levels 1, 2 and 3) reflect weaker performing schools (normally situated in rural or township areas, where no tuition fees are paid) indicating a lower socio-economic status.

In Table 1, we summarise the categorical covariates considered in this study. Most students are from a quintile 5 (well-performing) school, followed by the school quintile being unknown to the university. Students assigned to the unknown school quintile are students who completed their schooling outside the borders of South Africa (international students), where schools may not follow the same rating scheme, or perhaps no rating scheme.

Table 1. Descriptive statistics of the categorical covariates

Factor	Percentage
Gender	
Male	48%
Female	52%
Repeating the Module	
No	87%
Yes	13%
School Quintile	
1	4%
2	6%
3	10%
4	9%
5	39%
Unknown	32%

A good predictive model should be trustworthy and robust with as few covariates (also commonly referred to as independent or predictor variables) as possible. Bainbridge, Melitski, Zahradnik, Lauría, Jayaprakash and Baron (2015) studied some demographic, educational and behavioural patterns in search of finding promising covariates to predict at-risk students in an online Masters of Public Administration programme. Among their promising covariates were the number of times a student logged into their online course portal, and their participation in an online forum for the course. In summary, they revealed that combining these covariates with more traditional covariates, such as the gradebook, class size, and age, could enhance the predictive power of the model to identify at-risk students. Furthermore, special care must be taken when considering possible covariates for building a predictive model as different courses might require different information to train the model. For example, some courses

might require practical or laboratory work and others not, or a face-to-face course component compared to an online course component.

These unique differences require some strategic awareness from the educator in order to decide whether it is meaningful to incorporate these variables into the predictive model. Furthermore, it is of good practice and according to the traditional steps of a modelling cycle (compare COMAP-SIAM, 2016) to continuously update the training data and retrain the model to improve the robustness and predictive power of the model.

4. STATISTICAL ANALYSIS AND RESULTS

4.1 Predictive model

A number of forces should be considered when building a predictive model, such as, model simplicity, predictive power and the usability of the model in a course, where often these factors work against each other (Emmert-Streib & Dehmer, 2019). For example, to carry out a predictive model early on in an academic semester will allow sufficient time to assist at-risk students. However, this will come at an accuracy cost, as there will be less information available to 'train' the model. Furthermore, to increase the predictive power of a model might come at a simplicity and usability cost, as more predictor variables may be required to improve the accuracy. In this study, we followed the approach of finding a suitable balance between these opposing forces. The course outline for the Business Statistics course is displayed in Table 2. After careful consideration of the course outline, we decided to implement the predictive model in week 7 of the semester. We reason that this strategy should allow for a good balance between forces; it will allow educators enough time to assist the identified at-risk students without largely compromising the accuracy, and it will allow enough information (training data) to 'train' the model.

Table 2. Course outline of the Business Statistics course from week 1 - 14

Week	Quiz	Semester Test	Content
1			Sampling & Sampling Distribution
2 & 3	1		Confidence Intervals
4, 5 & 6	2		Hypothesis Testing: One Population Mean & Proportion
7		1 (week 7)	Revision
8	3		Hypothesis Testing: Chi-Square Procedures
9	4		Hypothesis Testing: ANOVA
10	5		Multiple Regression
11			Differentiation
12			Maxima and Minima
13		2 (week 13)	Revision
14			Linear Programing

The predictive power (trustworthiness) of the models used in this study was assessed by using the statistical calculations explained below (see Marbouti, Diefes-Dux & Madhavan, 2016):

$$Accuarcy = \frac{True\ Negatives + True\ Positives}{Total\ number\ of\ students}$$

$$Accuracy\ (Pass) = \frac{True\ Negatives}{Number\ of\ passed\ students} = \frac{True\ Negatives}{True\ Negatives + False\ Positives}$$

$$Accuracy\ (Fail) = \frac{True\ Positives}{Number\ of\ failed\ students} = \frac{True\ Positives}{True\ Positives + False\ Negatives}$$

$$F_{1.5} = \frac{(1 + 1.5^2) \times True\ Positives}{(1 + 1.5^2) \times True\ Positives + 1.5^2 \times False\ Negatives + False\ Positives}$$

Where:

- *true positives* denote the number of students that failed and were identified as at-risk,
- *true negatives* denote the number of students who passed the course and were not identified as at-risk,
- *false negatives* (also known as type II error) denote the number of students who failed the course but were not identified as at-risk students,
- *false positives* (also known as type I error) denote the number of students who passed the course but were identified as at-risk students, and
- $F_{1.5}$ denotes the harmonic mean of precision and recall. More specifically, the harmonic mean takes into account the accuracy for the students who passed and failed the course, where it weighs the accuracy for students who failed more than students who passed (Van Rijsbergen, 1979).

The traditional technique often used to identify at-risk students in statistics courses in the higher education context in SA (which is also the current technique used in this Business Statistics course), is to identify the students that obtain a mark less than 50% in their first formal assessment (semester test 1). Then, these students are categorised as at-risk of failing the course. This traditional technique will be referred to as the *baseline model* for this investigation, analysis and report findings. In Section 4.2.1, we illustrate that this model is not trustworthy in predicting at-risk students. Therefore, it seems necessary to find a better performing predictive model. Following this notion, we investigated four alternative predictive models in this study, namely:

- *Multiple Regression:* Multiple linear regression is an extension of simple linear regression by allowing for more than one independent variables.
- *Logistic Regression:* Logistic regression is a widely used prediction method in statistics and particularly in predicting at-risk students (see e.g. Bainbridge et al., 2015; Marbouti et al., 2016). Moreover, logistic regression calculates the likelihood of observing a binary variable (e.g. 0 – fail and 1 – pass), using multiple covariates (independent variables).

- *Decision trees:* A decision tree is a tree-like model that predicts responses by following the decisions in the tree from the root node to the leaf node. In this study we considered two trees described as
  i. a classification tree (the response variable is nominal, i.e. pass or fail), and
  ii. a regression tree (the response variable is numeric).

The advantages of the above models, compared to other predictive models found in the literature, are the ease in which these models can be implemented in the most basic statistical software packages. Moreover, when building a predictive model, one should keep in mind the following good statistical practices: (i) a good predictive model should be as powerful as possible with as few covariates as possible so that we are not overtraining the model. This occurs when the model maximises its performance on the training data by unknowingly modelling the residual noise as if it represented the underlying model structure, rather than learning to generalise from a trend; (ii) the selected covariates for a predictive model should be reliable and easily accessible to the educator. In addition, when deciding on a prediction model, one should also take into consideration the type of data used to train the model. For example, categorical covariates may yield better results in a classification tree than in a regression-based model.

4.2 Statistical analysis and results

Before building our predictive models, we started by assessing the causal relationship between the covariates and FM. In Table 3, we show the Pearson's correlation coefficient, which is a measure of the strength of the relationship between the students FM and the numeric covariates. Equivalent to the study of Marbouti et al. (2016), a Pearson's correlation coefficient of at least 0.3 is regarded as an acceptable covariate for further analysis, which was acceptable for all covariates in this study.

Table 3. Pearson correlation coefficients between covariates

Covariate	Pearson Correlation Coefficient
Semester test 1 mark	0.6197
Average quiz mark	0.4339
Mark for the prerequisite	0.3080

Table 4 shows the results for the chi-squared test for independence. This test determines whether the categorical covariates and FM (pass/fail) are statistically related. The only categorical covariate statistically related, at a 5% level of significance, to the FM is whether the student is repeating the course or not.

Table 4. Chi-squared independence test

Covariate	p-value
Repeating the module	0.0003
Gender	0.0751
School quintile	0.7530

In addition, Figure 1 (a)-(d) displays the relationship of each student's covariates to their course outcome (i.e. pass or fail). More specifically, Figure 1a shows the relationship between the prerequisite course mark (shown on the x-axis) and the semester test 1 mark (shown on the y-axis), Figure 1b shows the relationship between the average quiz mark and semester test 1 mark, Figure 1c shows the relationship between the prerequisite course mark and average quiz mark, and Figure 1d shows the relationship between all three numeric covariates.

To emphasize the challenging nature of building predictive models, it is meaningful to point out the irregular behaviour of some students. For example, in our dataset some students performed well in one or both of the covariates but failed the course. In contrast, some students passed the course but performed poorly in one or both of the covariates used. Similarly, Marbouti et al. (2016) highlighted that a students' behaviour is seldom the same throughout the semester. For example, in our study, many students did not write semester test two or stopped attending lectures due to financial constraints. Therefore, it seems unreasonable to expect that a model can precisely predict all the student's final outcome in the course – no model is faultless. Regardless of these difficulties, the traditional steps of the modelling cycle should be considered and continuous work should be carried out to improve the accuracy of the model over time.

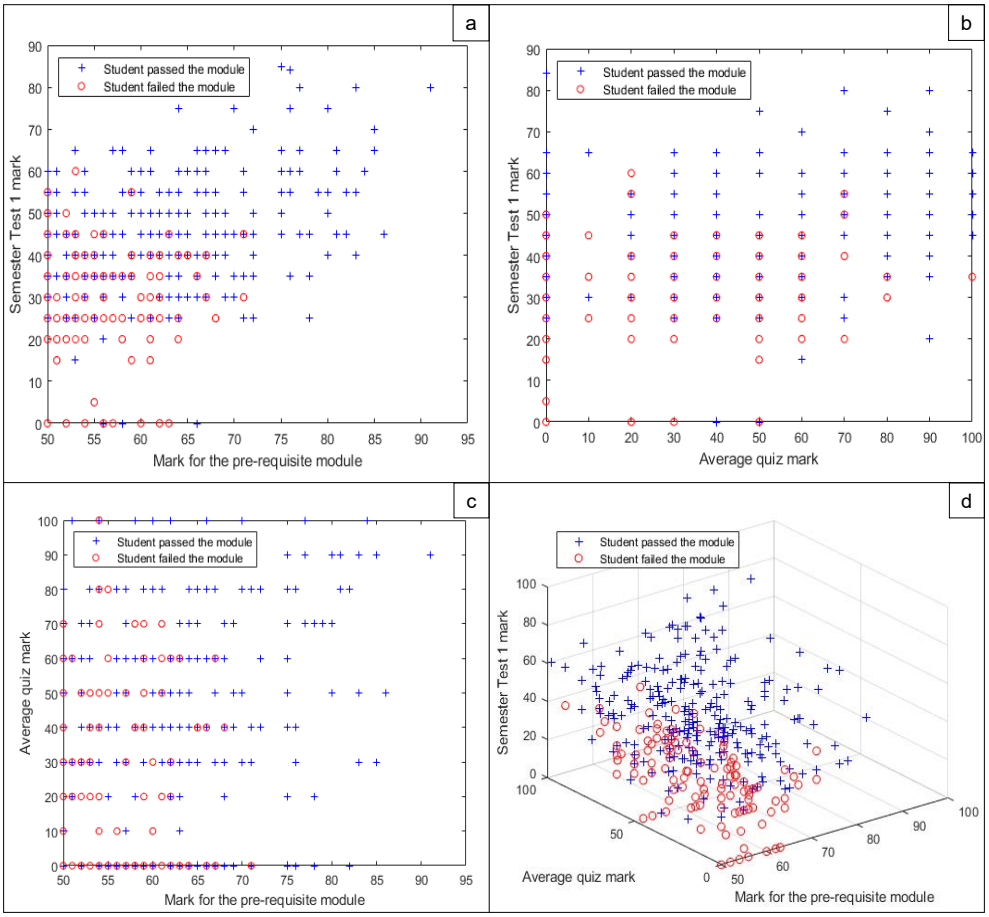

Figure 1. The relationship between the covariates and the module outcome

4.2.1 Model Validation

In order to satisfy the six evaluation criteria mentioned above, we start by partitioning the sample data into two datasets: 50% for training and 50% for model validation. This technique, known as cross-validation, assesses how the models will generalise to an independent dataset. The goal of the model validation is to assess how accurately a predictive model will perform in practice, by using the accuracy formulas above, and to prevent the likelihood of overfitting the model to the training set. To reduce variability in our analysis of the models, we implemented 1000 rounds of the cross-validation, using random samples, where the validation results are averaged giving us an estimate of the model's predictive performance. In addition, we also calculate the standard error (SE) of the accuracy estimates over the 1000 rounds. A good predictive model should yield high accuracy measures, a low number of *false positives (type I error)* and *false negatives (type II error)* with small SE measures. However, it is particularly important to have as few false negatives as possible as this classification carries a high consequence (i.e. to identify a student as not at-risk, but the student should have been identified as at-risk), where the *false positives* classification (i.e. to identify a student as being at-risk, when the student is not at-risk) carries less consequence. In keeping with the good statistical practices mentioned above, we found that the best covariates, based on this dataset, for the logistic regression model to be their semester test 1 mark and their prerequisite course mark, for the multiple regression model the average quiz mark, semester test 1 mark, and their prerequisite course mark. For the decision trees, we found the best covariates to be the average quiz mark, semester test 1 mark, their prerequisite course mark, and whether the student is repeating the course or not. Table 5 summarises the accuracy measures for the five predictive models used in this study.

Table 5. Measures of accuracy in the predictive models

Method	Base Method	Logistic Regression	Multiple Regression	Regression Tree	Classification Tree
$F_{1.5}$	43%	66%	72%	62%	61%
SE	4%	4%	3%	6%	6%
Accuracy	70%	80%	79%	75%	75%
SE	2%	2%	2%	3%	3%
Accuracy-Pass	86%	88%	80%	81%	82%
SE	2%	4%	3%	5%	5%
Accuracy-Fail	39%	65%	76%	62%	60%
SE	5%	6%	5%	8%	8%
True Negative*	57.7%	59.0%	53.6%	54.2%	54.7%
SE	2.5%	2.4%	2.6%	3.3%	3.1%
True Positive*	12.9%	20.2%	25.1%	20.6%	19.9%
SE	1.8%	2.0%	2.1%	3.0%	2.6%
False Negative*	20.2%	12.8%	8.0%	12.4%	13.1%
SE	2.0%	2.7%	2.0%	3.2%	3.1%
False Positive*	9.3%	8.0%	13.3%	12.7%	12.3%
SE	1.5%	2.3%	2.5%	3.1%	3.2%

*Represents the percentage of the sample.

From Table 5, the base model is not a suitable model for predicting at-risk students with a $F_{1.5}$ score of 43% and an overall accuracy of 70%. More importantly, it yielded the largest percentage of *false negatives.* Recall, the high consequence the false negative classification carries. Therefore, the base model is untrustworthy, as no intervention programme for at-risk students could be meaningful. The logistic regression and multiple regression models yielded far superior results with $F_{1.5}$ scores of 66% and 72%, respectively, and an overall accuracy of 80% and 79%, respectively. In addition, both models yielded a satisfactory percentage of *false negative* and *false positive* outcomes, making these two models superior to the base model. The regression and classification tree models did not perform as well (with a $F_{1.5}$ score of 62% and 61%, respectively and high SE) as the logistic and multiple regression models in this study. However, the decision trees may be more useful when using categorical variables that are related to the dependent variable. The expected percentage of the sample incorrectly predicted (*false negatives* + *false positives*) by the models are 29.5% for the base model, 20.8% for the logistic regression model, 21.3% for the multiple regression model, 25.1% for the regression tree, and 25.4% the classification tree. Thus, the logistic regression and multiple regression models yielded the most trustworthy results in identifying at-risk students, making these our models of choice in this study. In addition, the regression and decision tree models satisfied the six evaluation criteria outlined in Section 2.2 (compare Meyer, 2012). In particular, these models (i) yielded accurate results; (ii) are realistic and viable as the data required in the building of these models are relatively easy to obtain and implement in many statistical packages; (iii) yielded precise predictions (i.e. pass or fail); (iv) are known to be robust with extensive literature available to test the robustness of these models (although this is beyond the scope of this inquiry); (v) have successfully been used to predict at-risk students in this Business Statistics course, and (vi) yielded useful results, which will allow educators to identify at-risk students more accurately.

An area, worthwhile for further investigation would be to determine how well these predictive models perform in other courses and particularly in other mathematically related courses. Such an investigation might point to a model that is robust enough to give accurate results across many courses. Furthermore, it would also be worthwhile to consider incorporating a class attendance covariate into these models, especially, since we as educators mostly have the perception that academically successful students have good class attendance. It is often challenging to record data from class attendance, especially in large classes, where not all students have smart devices for an electronic attendance recording system, especially in a developing country such as South Africa. Another covariate to consider could be how active the students are on the electronic learning environment and how often they visit information and support material on this platform.

In addition, both the logistic regression and multiple regression models forecasted over 30% *(true positives + false positives)* of the students in the study as being classified as at-risk of failing the course. These large numbers could place enormous strain on educators to provide sufficient support for at-risk students, after all, it will be a fruitless exercise to identify students as at-risk, and not providing sufficient academic support.

5. CONCLUSION

In this study, we developed five different predictive mathematical models to identify at-risk students as early as possible in the academic semester in a Business Statistics course at a public university in SA. Quantitative and qualitative data were collected from past

Business Statistics students and a number of numerical and categorical covariates and their relationships were investigated to answer the two research questions: (1) what is a suitable predictive mathematical model used to identify at-risk students in a Business Statistics course at tertiary level, and (2) how effective is such a model to predict students' academic success in this course? We followed traditional modelling steps to construct the different predictive models (based method, logistic regression, multiple regression, regression tree, classification tree), commonly found in the literature, and compared the accuracy of these models based on Meyer's criteria (2012). A good selection technique for a predictive model should be an ideal balance between accuracy and simplicity. In particular, the logistic regression and multiple regression models yielded the most truthful results, using a cross-validation test. More specifically, these models yielded the highest accuracy with the smallest standard errors. An interpretation of model results might inform educators early in the academic semester of potential *at-risk* students, where educators will have the opportunity to intervene by providing rich academic support in the learning of statistics. Early detection of possible academic failure with suitable treatment can improve throughput rates in statistics courses without compromising academic standards.

Furthermore, our efforts of finding a suitable predictive model is aligned with the notion of learning mathematical and/or statistical knowledge by highlighting the different strands of Kilpatrick et al. (2001) - conceptual understanding, procedural fluency, strategic competence, adaptive reasoning and productive disposition. Apart from the knowledge base, to ultimately develop statistical literacy, reasoning and thinking, and an interconnection between these statistical components (compare Garfield & Ben-Zvi, 2007), a notion of supporting disposition is also present with the belief that every student can attain the necessary skills for academic success when formal education meets the students' needs and not the incapacity of students. Further research could follow; into efficient intervention programmes for at-risk students and at combining models to build a 'new' hybrid model to predict at-risk students more accurately.

REFERENCES

Bainbridge, J., Melitski, J., Zahradnik, A., Lauría, E. J.M., Jayaprakash, S. & Baron, J. 2015. Using learning analytics to predict at-risk students in online graduate public affairs and administration education. *Journal of Public Affairs Education*, 21(2), 247-262. Retrieved from http://www.jstor.org/stable/24369796. https://doi.org/10.1080/15236803.2015.12001831

Bernstein, A. 2013. Mathematics outcomes in South African schools. What are the facts? What should be done? *The Centre for Development and Enterprise*. South Africa. Retrieved from http://www.cde.org.za

Blum, W. & Leiβ, D. 2007. How do students and teachers deal with modelling problems? In C. Haines, P. Galbraith, W. Blum, & S Khan (Eds.), *Mathematical modelling: Education, engineering and economics* (pp. 222-231). Chichester: Harwood. https://doi.org/10.1533/9780857099419.5.221

Boaler, J. (Ed.). 2000. *Multiple perspectives on mathematics teaching and learning*. London: ABLEX Publishing.

Cassidy, S. 2015. Resilience building in students: the role of academic self-efficacy. *Frontiers in psychology*, 6,1781. https://doi: 10.3389/fpsyg.2015.01781

Coetzee, S. & van der Merwe, P. 2010. Industrial psychology students' attitudes towards statistics. *SA Journal of Industrial Psychology /SA Tydskrif vir Bedryfsielkunde*, 36(1), 1-8. https://doi.org/10.4102/sajip.v36i1.843

COMAP-SIAM. 2016. *Guidelines for assessment & instruction in mathematical modeling education* (GAIMME). Norman, Oklahoma. Available from http://www.siam.org

Doerr, H.M., Ärlebäck, J.B. & Misfeldt, M. 2017. Representations of modelling in mathematics education. In: GA Stillman, W Blum, & G Kaiser (Eds.), *Mathematical modelling and applications* (pp. 71-81). Cham: Springer. https://doi.org/10.1007/978-3-319-62968-1_6

Emmert-Streib, F. & Dehmer, M. 2019. Evaluation of Regression Models: Model Assessment, Model Selection and Generalization Error. *Mach. Learn. Knowl. Extr.*, 1, 521-551. https://doi.org/10.3390/make1010032

Garfield, J. & Ben-Zvi, D. 2007. How students learn statistics revisited: A current review of research on teaching and learning statistics. *International Statistical Review*, 75(3), 372-396. https://doi.org/10.1111/j.1751-5823.2007.00029.x

Greefrath, G., Koepf, W. & Neugebauer, C. 2017. Is there a link between Preparatory Course Attendance and Academic Success? A Case Study of Degree Programmes in Electrical Engineering and Computer Science. *International Journal of Research in Undergraduate Mathematics Education*, *3*(1), 143-167. https://doi.org/10.1007/s40753-016-0047-9

Kilpatrick, J., Swafford, J. & Findell, B. (Eds.) 2001. *Adding it up: Helping children learn mathematics*. Washington, DC: National Academies Press.

Lesh, R. & Doerr, H. 2003. Foundations of a models and a modelling perspective on mathematics teaching, learning, and problem solving. In R. Lesh & H. M. Doerr (Eds.), *Beyond constructivism: Models and modelling perspectives on mathematics problem solving, learning, and teaching* (pp. 3-33). Mahwah: Lawrence Erlbaum. https://doi.org/10.4324/9781410607713

Marbouti, F., Diefes-Dux, H.A. & Madhavan, K. 2016. Models for early prediction of at-risk students in a course using standards-based grading. *Computers & Education*, 103, 1-15. Retrieved from http://dx.doi.org/10.1016/j.compedu.2016.09.005

Meyer, W.J. 2012. *Concepts of mathematical modeling*. Mineola, New York: Dover Publications, INC.

Onwuegbuzie, A.J. 2004. Academic procrastination and statistics anxiety. *Assessment & Evaluation in Higher Education*, 29(1), 3-19. https://doi.org/10.1080/0260293042000160384

Rach, S. & Heinze, A. 2017. The transition from school to university in mathematics: Which influence do school-related variables have? *International journal of science and mathematics education*, 15, 1343-1363. https://doi.org/10.1007/s10763-016-9744-8

Samuelsson, J. 2010. The impact of teaching approaches on students' mathematical proficiency in Sweden. *International electronic journal of mathematics education*, 5(2), 61–78.

Slavin, R.E. & Madden, N.A. 1989. What works for students at risk: A research synthesis. *Educational leadership*, 46(5), 4-13.

Talsma, K., Schüz, B., Schwarzer, R. & Norris, K. 2018. I believe, therefore I achieve (and vice versa): A meta-analytic cross-lagged panel analysis of self-efficacy and academic performance. *Learning and Individual Differences*, 61, 136-150. https://doi.org/10.1016/j.lindif.2017.11.015

Van Appel, V. & Durandt, R. 2018. Dissimilarities in attitudes between students in service and mainstream courses towards statistics: an analysis conducted in a developing country. *EURASIA Journal of Mathematics, Science and Technology Education*, 14(8). https://doi.org/10.29333/ejmste/91912

Van Rijsbergen, C.J. 1979. Information retrieval (2nd ed.). London: Butterworths.

Van Zyl, A., Gravett, S. & De Bruin, G.P. 2012. To what extent do pre-entry attributes predict first year student academic performance in the South African context? *South African Journal of Higher Education*, 26(5), 1095-1111. https://doi.org/10.20853/26-5-210