AUT UNIVERSITY
Faculty of Business and Law

**Department of Economics**
*Working Paper Series*

# Using Predictive Modelling to Identify Students at Risk of Poor University Outcomes

Pengfei Jia and Tim Maloney

2014/03

# Using Predictive Modelling to Identify Students at Risk of Poor University Outcomes

Pengfei Jia and Tim Maloney*

April 2014

Keywords: Predictive Risk Modelling, University Failure and Dropout Behaviour, and New Zealand

JEL-Classifications: I21, I22 and I28

*Acknowledgement and Disclaimer: Access to the data used in this study was provided by a public university in New Zealand for the agreed purposes of this research project. The interpretations of the results presented in this study are those of the authors and do not reflect the views of this anonymous university.*

* Correspondence to: Tim Maloney, Economics Department, Auckland University of Technology, Private Bag 92006, Auckland 1142, New Zealand, (tim.maloney@aut.ac.nz).

**Abstract**

We use predictive modelling to identify students at risk of not completing their first-year courses and not returning to university in the second year. Our aim is two-fold. Firstly, we want to understand the pathways that lead to unsuccessful first-year experiences at university. Secondly, we want to develop simple, low-cost tools that would allow universities to identify and intervene on vulnerable students when they first arrive on campus. This is why we base our analysis on administrative data routinely collected as part of the enrolment process from a New Zealand university. We assess the 'target effectiveness' of our model from a number of perspectives. This approach is found to be substantially more predictive than a previously developed risk tool at this university. Students in the top decile of risk scores account for over 29% of first-year course non-completions and more than 23% of second-year student non-retentions at this university.

## 1. Introduction

Poor outcomes at university are a concern to students, institutions and public funding bodies (Grubb, 1989; Hartog et al., 1989; Tinto, 1993; and Montmarquette et al., 2001). This may be a by-product of the rapidly rising university participation rates in many countries over recent decades. Course non-completion and dropout rates may be increasing as less able or academically prepared students are admitted to these universities. Public funding authorities are also increasingly concerned by the potential waste of public expenditures on students who subsequently fail at university. For example, the reduction in non-completion rates is a core concern of recent reforms of the tertiary education sector in New Zealand. Universities are given specific targets to achieve in course completion rates in order to maintain their level of public funding (e.g., see New Zealand Ministry of Education, 2004).

There is a substantial body of empirical literature on the determinants of university non-completion outcomes (e.g., Wetzel et al., 1999; Montmarquette et al., 2001; Singell, 2004; Kerkvliet and Nowell, 2005; Bai and Maloney, 2006; Ishitani, 2006; Stratton et al., 2008; and Belloc et al., 2010). Although a comprehensive understanding of the relative importance of the various reasons for non-completion behaviour remains elusive, it has been widely recognized that individual characteristics, student educational backgrounds, and institutional factors are the main determinants of these outcomes (e.g., Robst et al., 1998; and Kerkvliet and Nowell, 2005). However, due mainly to limited data availability, most previous studies have utilized relatively few factors in their analysis. Using a more comprehensive dataset, our study is able to analyse the impact of a wide variety of explanatory variables on poor university outcomes.

Our paper uses an administrative dataset from a large public university in New Zealand to estimate the determinants of course non-completion in the first year and university non-retention in the second year. Administrative data have a number of advantages for the purposes of this study. Firstly, our database covers the entire first-year cohorts of students over four years (2009-2012). Secondly, these data are gathered as part of the normal application process, and thus no additional expense or

inconvenience is incurred in acquiring information (as would be needed in conducting a survey of first-year students). Thirdly, because these data are collected for enrolment purposes, the variables and their definitions are consistent over time. This is an important aspect if we want to use historical data to predict the at risk status of future students.

However, there are some disadvantages in using administrative data not specifically collected for the purposes of this research. Potentially important background factors, such as family income and expected financial support while studying at university, that have been widely discussed in the literature are unavailable in administrative database (e.g., Stampen and Cabrera, 1988; McPherson and Schapiro, 1991; DesJardins et al., 2002; Kerkvliet and Nowell, 2005; Ishitani, 2006; Montmarquette et al., 2007; and Stratton et al., 2008). To our knowledge, only a few studies have used administrative data to analyse poor university outcomes (Robst et al., 1998; Singell, 2004; Bai and Maloney, 2006; and Belloc et al. 2010).

This study has two goals. Firstly, we want to estimate the effects of a wide array of factors that may lead to both first-year course non-completion and second-year university non-retention outcomes. Secondly, we want to use these results to test the efficacy of a potential predictive risk tool for the early identification of students who are vulnerable to adverse outcomes at university. This is a trial to show how existing administrative data could be used in targeting intervention services (e.g., special tutorials, student advising and mentoring services) at the most vulnerable students entering university for the first time. We do this by randomly splitting our original sample in two. We use the first subsample to estimate maximum likelihood probit models on course non-completion and university non-retention outcomes. We then use the second subsample to predict these outcomes and compare them to the actual outcomes experienced by these students. This gives us an estimate of the 'target efficiency' of this predictive risk tool. We can also compare these results to an existing risk tool used by this university which was based on a survey of new entrants. We can then compare the target efficiency of our objectively constructed predictive risk tool using lower-cost existing administrative data with a subjectively constructed assessment tool based on higher-cost survey data.

The rest of the paper is organized as follows. Section 2 provides a brief overview of the relevant existing literature. Section 3 describes the data used in our regression and summarises our econometric approach. Section 4 analyses our empirical results. Finally, Section 5 draws conclusions from this analysis, and suggests possible directions for future work in this area.

## 2. Literature Review

Predictive Risk Models (PRMs) have been used previously in such areas as health care and child protection (e.g., see Billings et al., 2006, 2012; Vaithianathan et al., 2013). To our knowledge, this PRM approach has not been applied previously to the analysis of students at risk of adverse academic outcomes at university.

However, there is a substantial literature estimating the factors that influence poor student university experiences. For example, many studies have shown substantial differences in student demographic characteristics (e.g., ethnicity, gender, country of origin and age) associated with dropout behaviour (e.g., see Grayson, 1998; Robst et al., 1998; Wetzel et al., 1999; Montmarquette et al., 2001; Bai and Maloney, 2006; Mastekaasa and Smeby, 2008; Belloc et al., 2010; and Rodgers, 2013). For example, using administrative data from an urban university in New Zealand, Bai and Maloney (2006) found that the average dropout probabilities were 7.4 and 9.1 percentage points higher for Māori and Pacific Island students, respectively, compared to otherwise observationally equivalent students of other ethnicities. Similar evidence on ethnicity differences in retention rates were also provided by Wetzel et al. (1999) who analysed the factors affecting the dropout probabilities of students during their first two years of study at a U.S. university. In addition, other research has found that gender has a significant impact on the decision to drop out of university (Montmarquette et al., 2001, 2007; Mastekaasa and Smeby, 2008; and Belloc et al., 2010).

It has been widely recognized that student's success at university is substantially affected by his or her prior academic performance (e.g., see Betts and Morell, 1999; Cohn et al., 2004; Cyrenne and Chan, 2012; and Ficano, 2012). However, fewer studies have empirically examined the impact of past academic performance on student dropout behaviour (Wetzel et al., 1999; Montmarquette et al., 2001; Singell,

2004; Bai and Maloney, 2006; Ishitani, 2006; Stratton et al., 2008; Belloc et al., 2010; and Ost, 2010). For students new to the university, Singell (2004) found positive effects on university retention from the previous Grade Point Average (GPA) of the student. Montmarquette et al. (2001) and Bai and Maloney (2006) also found that early academic performance at university was a key determinant of subsequent dropout behaviour.

Institutional characteristics, such as class size and the overall difficulty of specific university programmes, could also play a role in student academic success (e.g., Tinto, 1982). Numerous studies have considered the importance of class size on high school academic performance (e.g., Angrist and Lavy, 1999; Krueger, 1999, 2003; Hoxby, 2000; Dobbelsteen et al., 2002; Rivkin et al., 2005). For example, according to Krueger (2003), the Tennessee Student/Teacher Achievement Ratio (STAR) experiment showed that students who were randomly assigned to small classes had better academic achievement outcomes than those placed in larger classes. To our knowledge, no previous published work has considered the effects of 'class size' on academic outcomes at university. We use non-experimental data in our study to estimate the effects of various aspects of class size on the probability of course non-completion and university non-retention.

Past research confirms the considerable differences of study areas on student dropout behaviour (e.g., Robst et al., 1998; Rodgers, 2013). Students who study science or engineering may be more likely to drop out than those who study arts or business, possibly due to the degree of difficulty of course material and academic expectations in these programmes. For example, using administrative data from State University of New York at Binghamton, Robst et al. (1998) found that students in the School of Management are more likely to be retained than students in the School of Arts and Sciences.

A number of previous studies have made more theoretical contributions in modelling student non-completion or dropout behaviour (e.g., see Altonji, 1993; Manski, 1989; Light, 1996; and Stinebrickner and Stinebrickner, 2012). For example, the student integration model by Tinto (1993) is one of the most comprehensive theoretical frameworks, in which he emphasizes the importance of academic and social

integration in predicting retention. Another seminal work by Bean (1980) incorporates external factors into the intention of students to either stay or leave university. In addition, DesJardins et al. (1999) applied an event history model in examining the temporal dimensions of student dropout behaviour. By developing a two-period model, Light and Strayer (2000) investigated whether the 'match' between student ability and college quality is an important determinant of university completion.

## 3. Data and Methodology

Administrative data used for this study were provided by a large public university in New Zealand. Data were made available on all first-year students who enrolled in Bachelor degree programmes at this university for the first time during the 2009 through 2012 academic years. The full sample contains 18,638 individuals and 101,948 course-specific observations. Individual student observations are used to examine non-retention outcomes in the second year, while individual course observations are used to investigate course non-completion outcomes in the first year.

Variable definitions and descriptive statistics are provided in Table 1. Our dataset contains detailed information typically available at the time of initial enrolment at university (e.g., year of entry, demographic characteristics, high school academic performance, and course and programme enrolment information).

(Insert Table 1 Here)

Two dependent variables are used in this study. These are course non-completion outcomes in the first-year and university non-retention outcomes in the second year.[1] The first dummy variable is set equal to one if the student did not successfully complete a course (i.e., receive a passing grade) in the first year; zero otherwise. The second dummy variable is set equal to one if the student did not return to re-enrol at

[1] We do not distinguish in this analysis between course dropouts (i.e., individuals who discontinued study prior to the end of the semester and dropped out of the course) and true fails (i.e., individuals who continued to the end of the semester, completed all assessments, but failed the course). This is largely because of the government reporting requirements in New Zealand that emphasise non-completion outcomes as result of either process.

this university at the beginning of the second year; zero otherwise. The results in Table 1 show that the mean course non-completion rate is 0.154 for the 101,948 course observations in our sample. The mean of non-retention rate is 0.226 for the 18,638 first-year student observations in our sample. Of course, students may leave university either temporarily or permanently, and for a variety of reasons.[2]

We have data on all first-year students from four annual cohorts (years 2009 through 2012). Our observations are fairly evenly distributed across these four cohorts (see Table 1). We include five dummy variables for a student's self-reported ethnicity (i.e., Asian, European, Māori, Pacifica, and other ethnicities). The latter is a residual category of all other reported ethnicities. The final category (Unknown) includes students who did report their ethnicity. As shown in Table 1, European ethnicities account for 39.2% of all first-year students at this university over this four-sample period. The second most common ethnic group is Asian (24.2%), following by Pacifica and Māori students (11.2% and 9.8%, respectively).

We use seven dummy variables on country of origin. Most first-year students are from New Zealand (69.5%), followed by China (8.6%), Korea (2.2%), India (1.6%) and Vietnam (1.3%). All other reported countries of origin are combined into a residual category 'Others' accounting for 15.5% of first-year students. Those not reporting their country of origin make up 1.3% of our sample. It is worth noting that ethnicity and country of origin could be very different in this research due to the fact that New Zealand has historically experienced a substantial inflow of migrants. For example, a student who reports 'Asian' as his or her ethnicity could also report 'New Zealand' as his or her country of origin.

Other personal characteristics include being female (60.1% of our sample) and enrolling for study part-time (29.3%). Just over one-half of our first-year students (57.8%) report information on their first or primary language. Of those who do, 70.7% identify English as their first language. Domestic students are defined as those

[2] Possible explanations for dropout behaviour include students struggling academically at university, transferring to other institutions, leaving for employment opportunities, etc. It should be noted that we have no information in the database on the reasons why individuals may have failed to return to this university at the beginning of the second year.

receiving domestic funding status (i.e., government subsidies). They comprise 87.9% of the first-year students at this university. The mean age for first-year students is 22.075. The average age might be higher than some other universities, mainly due to the high portion of part-time students at this institution.

Our dataset contains some information on the high school records of these students. Most high school students in New Zealand sit the National Certificate of Educational Achievement (NCEA) exams in the last three years at school.[3] These are national end-of-the-year exams across a number of compulsory and optional subject areas. Our dataset includes a summary measure of the overall performance on these NCEA exams in the final year of high school. As indicated in Table 1, this NCEA score is available for less than half of the first-year students across our cohorts ('Known NCEA Score'). We explain below how students can enter this university without NCEA results, but we also note that there may be some missing information on NCEA exams in the university's database. We will return to this issue in the recommendations for future research in this area in the concluding section of this study.

Two additional variables are available on the educational background of our students. We have access to a dummy variable on concerns over the literacy or numeracy levels for these individuals. A value of one for this variable ('Literacy/Numeracy') indicates that a test was taken during high school to investigate possible issues over appropriate literacy and numeracy levels. This test was taken by nearly one-quarter (23.8%) of the students in our sample. Finally, we know the identities of high schools in which these students were most recently enrolled. In New Zealand, these schools are sorted into deciles based on the socio-economic status of residents in the school catchment area.[4] For example, a decile 1 high school is among the 10% of schools from poorest socio-economic areas, while a decile 10 high school is among 10% of schools from the wealthiest socio-economy areas. The mean school decile in our sample is 6.846,

[3] For more information on the NCEA system see http://www.nzqa.govt.nz/qualifications-standards/qualifications/ncea/.

[4] For more information on the process used to determine the calculation of school deciles see http://www.minedu.govt.nz/NZEducation/EducationPolicies/Schools/SchoolOperations/Resourcing/OperationalFunding/Deciles/HowTheDecileIsCalculated.aspx.

indicating that these first-year students are drawn predominantly from higher decile schools.

There are six specified ways in which these first-year students could be granted entry to this university. The most conventional entrance type is 'NCEA Admission'. More than one-third of first-year students in our sample (36.3%) gain admission to this university through their NCEA scores.[5] Other students enter through 'Special Admission' status which refers to entering students who did not meet the NCEA entrance requirements, but entered because their other experiences (i.e., they had reached age 21 or above). Special Admissions account for 13.0% of first-year students in our sample. The variables 'Internal' and 'External' refer to students who gained university entrance because of previous study at this or another university. Internal entrants (8.9%) had completed a 'pre-degree' certificate or diploma at this particular university. External entrants (15.0%) had previously attended another university. A few students enter this university through the completion of Cambridge or International Baccalaureate programmes at secondary school ('Cambridge/IB'). Relatively few students enter through this more prestigious and challenging programmes (1.4%). These are typically high-achieving students from private high schools, who can use these secondary school qualifications to apply for universities both inside and outside of New Zealand. Finally, there are a number of other ways in which students can gain entry to this university. These are included in the residual category 'Others', and comprise slightly more than one-quarter of all first-year students in our sample (25.3%). Primarily, these include foreign students who gain entry through equivalent overseas high school qualifications.

For the purpose of analysing course non-completion outcomes, our administrative dataset contains some potentially useful information on the characteristics of these courses. We know from the recorded information at the outset of the academic year, the recommended 'Study Hours' in a course over the semester. Most courses (84.4%) report 'Known Contact' hours. These include scheduled lecture, tutorial, workshop and lab hours. They could also include scheduled office hours, and group study hours and generic academic preparation workshops in areas such as English, writing skills

[5] This also includes a few students who entered through their Bursary or University Entrance qualifications. These qualifications were replaced by the NCEA system in 2004.

and mathematics. We also know the average 'Class Size' and the 'Course Size'. The latter is the total number of students enrolled in the course. The former is the average number of students in a classroom. For example, a large first-year course could have 1,000 students enrolled. This would be the course size. These students could be taught in a single large class of 1,000, or they could be taught in 20 classes with an average class size of 50 students. We consider the separate effects of both course and class size on course non-completion outcomes. We also know whether or not the course is supported with internet content. Finally, we know the academic level of the course. A 'Level 4' course contains content that is intended for students below a Baccalaureate degree level. Most courses taken by these first-year students (83.6%) are intended for the first year of university study ('Level 5'), but some students enrol in courses intended for second and third-year study ('Level 6' (15.6%) and 'Level 7' (0.4%), respectively).[6]

We have additional academic information on students including the number of courses in which they have enrolled ('Courses Taken') and the proportion taken at Levels 6 or 7 ('High Level'). We also know whether or not the student has enrolled for a double-degree programme, and whether or not study is relegated to a single campus at this university. Less than 1% of first-year students enrol for a double degree, partly due to stringency of the entry requirements.

Finally, our dataset contains information on the initial programme of study. We use dummy variables to identify the largest 11 Bachelor degree programmes. The residual category includes all of the smaller degree programmes (7.5% of students in our sample). The largest three programmes are the Bachelor of Business (28.2%), the Bachelor of Health Sciences (19.5%), and the Bachelor of Arts (7.8%).

Maximum likelihood probit analysis will be used to estimate the effects of these various individual, school and enrolment factors on our two dichotomous dependent variables. The basic probit model can be written:

$$Y_i^* = \boldsymbol{\beta X}_i + u_i \qquad (1)$$

[6] The typical university baccalaureate programme in New Zealand is completed in three years of full-time study.

where $Y_i^*$ is a latent variable associated with course non-completion or student non-retention propensities. What we observe is a dummy variable $Y_i$ that equals 1 if the course was not completed in the first year, or the student did not return to re-enrol at this university in the second year; 0 otherwise. This depends on the latent dependent variable crossing an arbitrary threshold of zero.

$$Y_i = \begin{cases} 1 & if\ Y_i^* > 0 \\ 0 & if\ Y_i^* \leq 0 \end{cases} \qquad (2)$$

All of the aforementioned factors are included in the vector $\boldsymbol{X_i}$. The unknown coefficients are represented by the $\boldsymbol{\beta}$ vector which will need to be estimated. Finally, the random disturbance $u_i$ is assumed to have a normal, i.i.d. distribution. The probability of course non-completion or student non-retention can be denoted as the following:

$$P(Y_i^* > 0) = P(\boldsymbol{\beta X_i} + u_i > 0) = P(u_i > -\boldsymbol{\beta X_i}) = \boldsymbol{\Phi}(\boldsymbol{\beta X_i}) \qquad (3)$$

where $\boldsymbol{\Phi}(\cdot)$ is the Cumulative Distribution Function (CDF) of the standard normal.

We will use the average marginal effects to describe the influence of a one-unit change in an explanatory variable on the probability of course non-completion or student non-retention. This is because Probit model is a non-linear function of the coefficients, and the marginal effects are dependent on the values of the other repressors. For any particular factor $X_k$, this partial derivative can be written:

$$\frac{\partial P(Y_i = 1|\boldsymbol{X_i})}{\partial X_k} = \beta_k \phi(\boldsymbol{\beta X_i}) \qquad (4)$$

where $\phi(\cdot)$ is the Probability Distribution Function (PDF) of the standard normal.

This probit estimation is also used in the development of our Predictive Risk Models (PRMs). The PRMs are designed to generate risk scores for any first-year student enrolling at this university. To assess the effectiveness of these predictive risk tools,

we compare our predicted outcomes against the actual outcomes for each of our dependent variables. For this reason, our full sample is randomly split into two equal-sized 'estimation' and 'validation' samples (e.g., see medical applications of this methodology in Billings et al., 2012). The estimation samples will be used to estimate the probit models, and the validation samples will be used to assess how well the PRMs correctly identify the actual course non-completion outcomes and student non-retention outcomes.

PRM performance is often summarised by reporting the area under the Receiver Operator Characteristic (ROC) curve. The ROC curve characterizes the relationship between sensitivity and specificity. It shows the trade-off between true positives (sensitivity) and false negatives (1-specificity). In the context of this research, for example, sensitivity is the probability that course non-completion is correctly identified by the model. Specificity is the probability that course non-completion outcome is incorrectly identified.

The area under the ROC curve measures how well the PRM accurately distinguishes course completion and non-completion outcomes. The larger the area under the curve, the better the PRM is at assessing risk. A model with 100% area under the ROC curve is said to have perfect fit, while a model with 50% area under the ROC curve is no better than tossing a coin in predicting whether or not a course will be successfully completed.

Another approach to assess the predictive power of our PRM is to compare the estimated at risk scores based on the regression analysis to the actual observed outcomes on course non-completion or student non-retention. Once we have generated risk scores for all the observations in our validation samples, we will sort these predicted risk scores into deciles and compare the predicted and actual outcomes. For example, for the course enrolments with the 10% highest risk scores predicted by our model, what proportion of actual course non-completions are captured? This simple comparison provides a useful indicator of the 'target effectiveness' of our predictive risk tool.

## 4. Empirical Results

Two separate maximum likelihood probit models were estimated for this study, one for course non-completion outcomes in the first year, and another for student non-retention outcomes in the second year. As noted above, we randomly selected 50% of our full sample for this estimation, and used the remaining 50% of the sample to validate our predictive risk models.[7] Because the probit models are nonlinear in the parameters, estimated coefficients do not have the usual linear least-squares interpretations (i.e., they do not measure the change in the probability of non-completion or non-retention given a one-unit change in the explanatory variables). Therefore, the marginal effects in this analysis are computed as sample mean of the individual marginal effects.[8] The estimated coefficients, standard errors, and average marginal effects are presented in Table 2 for both dependent variables.

(Insert Table 2 Here)

### 4.1 Results on First-Year Course Non-Completion

The descriptive statistics in Table 1 showed that the mean course non-completion rate was 15.4% for all first-year students in our sample over the period 2009 to 2012. The estimated determinants of this probability of course non-completion are reported in the first three columns of Table 2 for the 50,932 first-year course outcomes in our estimation sample. Holding constant other measured factors, we find that the probability of course non-completion varies systematically across the years (where 2012 is the excluded or benchmark category). All three estimated coefficients on the included year dummies are negative and statistically different from zero at better than a 1% level. This says that the course non-completion probability was highest in 2012 compared to the previous three years. However, given the lack of any clear time trend in these estimated marginal effects, it would be premature to conclude that these results suggest a systematic increase in course non-completion rates over time.

[7] Because we randomly selected 50% of the observations, the estimated coefficients and average marginal effects should be similar to the estimates for the full sample.
[8] The marginal effects could also be calculated at the sample means for the explanatory variables. For continuous functions in large samples, this technique yields similar results to the sample mean for the individual marginal effects.

Ethnicity appears to have a substantial impact on the probability of course non-completion. Relative to the omitted category ('Unknown' ethnicity), reporting being either European or Asian has a statistically significant negative impact on course non-completion outcomes. The estimated partial derivatives indicate that these mean reductions in the probability of course non-completion are -3.76 percentage points for Europeans and -2.60 percentage points for Asians. To assess the magnitudes of these effects, we can compare these estimates to overall sample mean. These are approximately 24.4% and 16.9% reductions in the average rate of course non-completion in our sample for these two ethnic groups, respectively. Reporting being either Pacifica or Māori has a statistically significant positive impact on course non-completion (these partial derivatives are 6.67 and 3.42 percentage points, respectively). This suggests that the difference between the estimated probability of course non-completion for an otherwise observationally equivalent first-year Pacifica student is approximately 10.43 percentage points higher than that of a European student. This is roughly two-thirds of the sample mean for this outcome.

The estimated results on country of origin require some explanation. The estimated coefficients on all six dummy variables are negative and statistically significant at better than a 1% level, compared to those who did not report their country of origin (i.e., the omitted category). In other words, those not reporting a country of origin at the time of initial enrolment at this university appear to be the highest risk group.

Consistent with earlier studies in this literature, female students have a lower estimated probability of course non-completion compare to male students. Holding other things constant, being female lowers this probability of course non-completion by 2.73 percentage points. This effect is statistically significant at better than 1% level. This gender effect is substantial in magnitude. It's equivalent to 17.7% of the mean rate of course completion in our sample.

Studying part-time study is estimated to substantially increase the rate of course completion in the first year. Being a part-time student increases this probability by 15.49 percentage points. This estimated marginal effect is slightly larger than the sample mean for this outcome. English as the first language has no measurable impact on course non-completion among the first-year in our sample. Being a

domestic student, surprisingly, increases the probability of course non-completion by 3.80 percentage points.

We use a series of dummy variables to allow for flexibility in the age effects on course non-completion. One dummy variable is used for being under the age of 18. A series of eight dummies are used for individual ages from 18 through 25 inclusive, and three dummies are used for age ranges 26 through 30, 31 through 35 and 36 through 45. The omitted age group is 46 and older. It's easy to characterise these findings. All individual ages from 18 through 25 have positive and significant effects on the probability of course non-completion relative to other age groups. Moreover, the results suggest that first-year students aged 20 or 21 are at the highest risk. Their probabilities of course non-completion are estimated to be, respectively, 7.13 and 7.02 percentage points higher than those of students aged 45 or older. It is worth noting that being either younger or older than these two ages steadily reduces the risk of course non-completion.

We find that students who scored higher on their NCEA exams in high school tend to have lower probabilities of course non-completion during their first year at university. Two variables must be considered in interpreting these results. The first is a dummy variable on having information on these exam results ('Known NCEA Score'), and the second is the composite exam score from this last year of high school ('Actual NCEA Score'). The estimated effect of having an NCEA score on this probability of course completion could be written:

$$\frac{\partial P(Non-Completion)}{\partial Known\ NCEA\ Score} = 7.60\% - 0.10\% \cdot Actual\ NCEA\ Score \qquad (5)$$

We know from Table 1 that the sample mean for those reporting a NCEA score is 155.107. Thus, for the average student with NCEA results, these exams reduce the probability of course non-completion in the first year by an average of nearly 8 percentage points:

$$\frac{\partial P(Non-Completion)}{\partial Known\ NCEA\ Score} = 7.60\% - 0.10\% \cdot 155.107 \approx -7.911\% \qquad (6)$$

Every 10-point increase in this NCEA score reduces the probability of course non-completion by a full percentage point.

The previous section indicated that the dummy variable on 'Literacy/Numeracy' tests in high school picks up possible concerns over the reading, writing and mathematics skills for students. As expected, taking these tests is associated with a significant increase in the probability of course non-completion in the first year at university. On average, it increases this probability of non-completion by 1.68 percentage points.

We expected that students from lower school deciles would have higher probabilities of paper non-completion during the first year at university. This result is largely confirmed by our analysis, but some discussion around these findings is needed. Firstly, the omitted category includes (mostly overseas) students who did not come from a high school with a decile ranking. All ten of the school deciles have positive and statistically significant effects on the probability of course non-completion in the first year at university. This again indicates that domestic students are generally at higher risk compared to international students. Secondly, although the largest estimated effects are found for the bottom four deciles, there is no evidence in these results that students from the top decile schools are at the lowest risk of course non-completion. In fact, there is some evidence of a 'U-shaped' relationship. The estimated effects for deciles 8 through 10 are positive and larger in magnitude than deciles 5 through 7. One possible explanation for these results is that many first-year students at this university coming from the highest decile schools were unable to gain admittance to higher ranked universities in New Zealand and overseas and generally did not have the same academic preparation (or motivation) of students from mid-decile schools.

The entrance types for students have measurable impacts on the probability of course non-completion in the first year. Students entering with a Cambridge or International Baccalaureate qualification are at the lowest risk. This entrance type reduces the probability of course non-completion by an average of 6.44 percentage points. Students with an 'External' entry (i.e., previous study at another university) are next lowest risk group. This entry type reduces the probability of course non-completion

by an average of 3.17 percentage points. The estimated effect on 'Internal' entry (i.e., holding a pre-degree qualification from this university) has no measurable impact on course non-completion (relative to the omitted category of other, unspecified entrance types). The two highest risk entrance types are 'Special Admission' and 'NCEA Admission'. The NCEA Admission standard is closely connected to the effect of NCEA Score discussed earlier. This isn't likely to be an at risk indicator because the vast majority of students entering on this basis will have reported a NCEA score that reduces the probability of course non-completion. However, the Special Admission standard is likely to be an indicator of vulnerability, because these are generally students in their twenties who lack school qualifications and enter on the basis of their age and work experience. This Special Admissions effect combined with at risk nature of students aged in their early twenties makes this a particularly vulnerable group.

The remaining covariates in this regression model relate to the courses or degree programmes in which these students were enrolled during their first year at university. We first ask whether course characteristics themselves play any role in the likelihood of course non-completion. One particular set of results is worthy of discussion here. We draw a distinction between the overall number of students enrolled in a course ('Course Size') and the average number of students in a classroom ('Class Size'). To ease the interpretation of the estimated results, both variables are divided by 10. There is a substantial literature on the effects of class size on student achievement at school. We extend this analysis to academic outcomes at university. However, there is one additional issue when considering university study. Individuals often enrol in large first-year courses, but these can be taught in either large settings (e.g., a single mass lecture) or small settings (e.g., multiple streams taught in smaller classrooms). These course and class size effects could be quite different for the probability of course non-completion. For example, courses with large enrolments could reduce the probability of course non-completion because of the introductory nature of the subject material and the need for large-scale assessment events. On the other hand, similar to the usual justification in the school literature, class settings with large enrolments could increase the probability of course non-completion due to the difficulty of students getting the individual attention they might need. These are precisely the direction of the effects that we find in our analysis. The estimated course size effect

is negative and statistically significant at better than a 1% level. We find that an increase in enrolment of 10 students in a course, on average, reduces the probability of course non-completion by 0.03 percentage points. The estimated class size effect is positive and statistically significant at better than a 5% level. We find that an increase in enrolment of 10 students in a class, on average, increases the probability of course non-completion by 0.10 percentage points. This suggests that course non-completion rates would be lowered by enrolling students in large first-year courses, but actually teaching them in smaller classroom settings.

Finally, our results indicate that programme study areas play an important role in course non-completion outcomes in the first year. We know the degree programmes in which these students initially enrolled.[9] The omitted category is the 7.5% of students who were enrolled in a number of relatively smaller programmes. Relative to this reference group, three programmes had significantly higher estimated rates of course non-completion. They are, in order of the size of these positive marginal effects, Bachelor of Mathematical Science (BMS 3.62 percentage points), Bachelor of Engineering Technology (MEngT 1.84 percentage points), and Bachelor of Arts (BA 1.72 percentage points). Seven programmes had significantly lower estimated rates of course non-completion. In order of the size of these negative marginal effects, the largest three programmes were Bachelor of Education (BEdu -10.89 percentage points), Bachelor of Design (BDe -6.79 percentage points), and Bachelor of Communication Studies (BCS -6.10 percentage points). Some caution should be exercised in interpreting these results. They could indicate something about the rigour or difficulty of first-year study in these areas, but they could equally indicate something about the unobserved characteristics of the students who enrol in these degree programmes.

### 4.2 Results on Second-Year University Non-Retention

The descriptive statistics in Table 1 showed that the mean non-retention rate for students was 22.6% in the second year at this university. The estimated determinants of this probability of non-retention are reported in the last three columns of Table 2 for the 9,301 students in our estimation sample. Holding constant other measured

[9] A student could be enrolled in more than one programme if they were doing a double degree,

factors, we find that the non-retention probability was significantly higher in the 2010 and 2011 cohorts when compared to the omitted 2012 cohort. However, given the fact that we have only four cohorts in this dataset, it would be premature to conclude that these results suggest a long-term decline in this second-year non-retention rate at this university.

Recall that both Pacifica and Māori students were significantly more likely to not complete their first-year courses compared to other ethnic groups. The only ethnic group with a statistically significant effect on non-retention in the second year is Māori. Specifically, Māori students have a probability of non-retention that is, on average, 5.85 percentage points higher than students without a reported ethnicity. This suggests that Pacifica students are the most likely among the reported ethnic groups to not complete their courses in the first year, while Māori students are the most likely ethnic group to not return to the university in the second year.

All of the marginal effects for student non-retention on the country of origin variables have the same negative and significant signs that they had for course non-completion. These combined results suggest that students not reporting a country of origin are at the highest risk of both course non-completion and student non-retention.

Female students are at lower risk of both course non-completion and non-retention. However, the latter effect is smaller in magnitude and weaker in statistical significance. Part-time students are substantially more likely to drop out of university in the second year. Studying part-time increases the probability of non-retention by an average of 15.80 percentage points. Thus, part-time study is arguably the single most important single at risk factor for poor university outcomes.

Domestic students are relatively more likely to discontinue their study at this university in the second year. This estimated average marginal effect of 6.74 percentage points is even larger in magnitude than the 3.80 percentage points we had found earlier for course non-completion. It is interesting that although age seemed to play an important role in course completion outcomes in the first year, it has no measurable effect on the probability of non-retention in the second year.

Students who scored higher on their NCEA exams in high school are less likely to drop out of university in the second year. Again, we need to estimate the marginal effect of having an NCEA score on this probability of non-retention using the two estimated average marginal effects:

$$\frac{\partial P(Non-Retention)}{\partial Known\ NCEA\ Score} = 5.30\% - 0.06\% \cdot Actual\ NCEA\ Score \qquad (7)$$

Using the sample mean from Table 1 for those reporting a NCEA score, we estimate that for those reporting these exam results, they reduce the probability of student non-retention by over 4 percentage points.

$$\frac{\partial P(Non-Retention)}{\partial Known\ NCEA\ Score} = 5.30\% - 0.06\% \cdot 155.107 \approx -4.006\% \qquad (8)$$

Every 10–point increase in this NCEA score reduces the probability of university non-retention by 0.6 percentage points.

The dummy variable on a 'Literacy/Numeracy' tests in high school again indicate a concern in these areas. These tests are associated with a significant increase on 3.74 percentage points in the probability of non-retention in the second year at this university. Recall that (largely domestic) students reporting a school decile were more likely to not complete their first-year courses. It is noteworthy that the same result does not hold for non-retention. None of the estimated coefficients on school deciles are positive and significant. In fact, students coming from schools in declies 4, 6, 7 and 10 are significantly less likely to be university dropouts in the second year.

Students entering this university with a Cambridge or International Baccalaureate qualification are both more likely to complete their first-year courses and to be retained by the university in the second year. Both 'External' and 'Internal' entry reduce the probability of non-retention in the second year. 'Special Admission' status, which was an at risk factor for course non-completion, has no measurable effect on non-retention.

We found earlier that students studying part-time are one of the most vulnerable groups for both course non-completion and university non-retention. In a similar way, enrolling in 6 or more courses and in a double degree both substantially reduce the probability of non-retention in the second year. These average estimated marginal effects are -20.81 and -13.11 percentage points, respectively. These effects may be related to unobservable personal characterises that lead students to enrol in 6 or more courses in the first year and allow them entry into double degree programmes.

Finally, consider the results on programme study areas for the non-retention outcomes for these students in their second year. Of the three highest-risk programmes for course non-completion, only the Bachelor of Arts degree has a significantly positive estimated coefficient on non-retention. The average marginal effect for the BA is a risk-increasing 8.73 percentage points. Of the three lowest-risk programmes for course non-completion, two of them have significantly negative estimated coefficients on non-retention. These are the Bachelor of Education (BEdu -5.94 percentage points) and the Bachelor of Communication Studies (BCS -4.75 percentage points). Thus, the degree programme with the lowest risk of course non-completion also has the lowest risk of university non-retention (the Bachelor of Education). The only other degree programme with a significantly lower non-retention outcome is the Bachelor of Business (BBus -3.30 percentage points).

### 4.3 Assessing the Predictive Power of Our PRMs

One way to assess the overall performance of our probit regression models is to consider the Pseudo $R^2$ statistics reported at the bottom of Table 2. The usual interpretation is that our models can explain approximately 13.19% of the variation in course non-completion outcomes in the first year and 10.63% of the variation in university non-retention outcomes in the second year within our estimation samples, respectively. These statistics, of course, only summarise the predictive power of our analysis *within* these samples. We want to know how well these models perform in predicting these outcomes *outside* of these estimation samples.

As noted earlier, we randomly selected 50% of our overall samples for estimation purposes, and retained the remaining 50% for assessing the performance of our Predictive Risk Models (PRMs). We report the area under the Receiver Operator Characteristic (ROC) curves for both course non-completion and student non-retention outcomes in the summary statistics of Table 2 using these respective validation samples. The ROC curves characterise the relationship between the 'sensitivity' and 'specificity' in these two models. Sensitivity is the probability that a course failure (or student dropout) outcomes is correctly identified. Specificity is the probability that a course completion (or student retention) outcome is correctly identified. The areas under these ROC curves indicate how well our PRMs perform in distinguishing between the respective outcomes in both areas. We can graphically illustrate the trade-offs between sensitivity and 1 - specificity at all possible thresholds. These results are shown in Figures 1 and 2.

(Insert Figures 1 and 2 Here)

The area under the ROC curve for course non-completion is 0.7553. This indicates that there is a 75.53% probability that a randomly selected course observation with a non-completion outcome will receive a higher risk score from our PRM than a randomly selected course observation with a completion outcome. This is an indicator of the 'target effectiveness' of this predictive risk tool could be compared to the results from other types of analyses (e.g., see Billings et al., 2006, 2012; and Vaithianathan et al., 2013). Similar interpretations can be given for the non-retention analysis with the area under ROC curve at 0.7125.

Another approach is assessing the effectiveness of our PRMs is to compare predicted to actual outcomes. We can use the regression results reported in Table 2 to compute risk scores for all course non-completion and student non-retention outcomes in our validation samples. We can then rank these predicted probabilities and sort them into deciles, and determine the proportion of actual adverse outcomes that we would capture with this procedure at every decile. Suppose we wanted to intervene (i.e., provide specific services) to students in the top decile (i.e., those with the highest 10% of risk scores). If our models were completely ineffective at predicting these outcomes, then the top 10% of risk scores would account for only 10% of the actual

adverse outcomes. The results in Table 3 indicate that the highest 10% of risk scores in the validation samples would capture 29.25% of course non-completions in the first year and 23.33% of student non-retentions in the second year. If we increased our delivery of interventions to the top two deciles, we would capture 47.57% of course non-completions and 40.91% of student non-retentions.

(Insert Table 3 Here)

It's generally difficult to provide any meaningful comparisons to the predictive power analysis of any particular PRM. Fortunately, in this situation we had information on an existing risk analysis tool developed by this university, which provides a convenient benchmark. The university had previously used the results from a survey administered to first-year students to predict who would likely experience academic difficulties over the first year of university study. Students were asked about their academic backgrounds, as well as their personal characteristics and views on university study. University administrators then attached 'weights' to these survey responses. These weights were based on subjective assessments on the relative importance of these various responses for course non-completions. No attempts had been made to base these weights on any type of objective analysis of this outcome of interest.

We constructed the risk scores from our validation sample using this administrative tool. We can then directly compare these predicted outcomes to the actual course non-completion outcomes. By any measure, the predictive power of this administrative tool was substantially inferior to our PRM. Because of 'ties' in adding up these risk measures using the integer weights, we can't select the highest 10% of risk scores. The approximate 'top decile' using the university's administrative tool accounted for 11.78% of course outcomes. These course outcomes accounted for 23.51% of course non-completions in this validation sample. The top decile of high-risk course outcomes using our PRM was nearly three-times more likely to experience a course non-completion than the overall sample (29.25/10.00). The top decile of high-risk course outcomes using the administrative tool was less than two-times more likely to experience a course non-completion (23.51/11.78). In this sense, our PRM was 46.5% more 'target effective' than the existing administrative tool.

The same comparisons can be made for the top two deciles using these two alternative approaches. Again, because of ties, the 'top two deciles' using the administrative tool accounted for 25.27% of course outcomes. These course outcomes accounted for 39.11% of course non-completions in this validation sample. The top two deciles of high-risk course outcomes using our PRM was 2.375-times more likely to experience a course non-completion than the overall sample (47.57/20.00). The top two deciles of high-risk scores using the administrative tool was 1.548-times more likely to experience a course non-completion (39.11/25.27). In this sense, the 'hit rate' of our PRM is approximately 53.4% higher than the existing administrative tool.

This relatively better performance of our PRM is not that surprising given that the administrative tool used by the university had never been appropriately validated. Moreover, this PRM approach has a very important additional advantage. The survey-based administrative tool requires the dissemination and processing of a first-year student survey each year. This can be an expensive operation. Our PRM tool, on the other hand, is based entirely on conventional and routine data collected as part of the enrolment process. Thus, once developed, there is virtually no additional on-going cost in using this PRM approach. In this sense, it is relatively more 'target effective' and 'cost efficient'.

## 5. Conclusion

This study has empirically estimated the determinants of course non-completion outcomes in the first year and student non-retention outcomes in the second year using administrative data from a large public university in New Zealand. These Predictive Risk Models have been developed to improve our understanding of the specific factors that that place students at risk of adverse outcomes early in their university careers. In addition, these PRMs could be potentially used by universities to develop effective, low-cost tools for identifying students at risk of adverse outcomes when they first arrive on campus. Such tools could be instrumental in delivering early interventions to those students most likely struggle at university.

Two dummy dependent variables were used in our regression analysis. These were course non-completion outcomes in the first year and student non-retention outcomes in the second year. Administrative data on courses and students were taken from four annual cohorts of students entering degree programmes for the first time at this university. Our findings suggest that a wide array of factors influence the probabilities of course non-completion and student non-retention. For example, part-time study is estimated to substantially raise the probabilities of both detrimental outcomes. Pacifica students are the ethnic group most at risk of course non-completion outcomes in the first year, while Māori students are the ethnic group most at risk of non-retention outcomes in the second year. Course non-completion rates and student non-retention rates are found to be significantly lower for female students. Better results on national high school exams substantially reduce the risk of non-completion and non-retention. Larger overall course enrolments are associated with lower course non-completion rates, but larger average class sizes are associated with higher course non-completion rates. Finally, both course non-completion and student non-retention outcomes vary substantially across the various degree programmes.

Our overall samples were randomly spilt into 'estimation' and 'validation' samples. The areas under ROC curves were 0.7553 and 0.7125, respectively, for the course non-completion and student non-retention outcomes. We found that the top risk decile of course observations could account for 29.25% of actual course non-completion outcomes in this validation sample. The top risk decile of student observations could account for 23.33% of actual student non-retention outcomes in this validation sample. These results are far better than the expected 10% of both outcomes that would be expected of a completely uninformative risk tool. These results were also better than the existing administrative tool used by this university to identify student at risk of course non-completions. We find that our PRM is at least 46.5% more target effective in identifying students vulnerable for course non-completions. We claim that out PRM would also be far more cost-effective because it would be based on existing administrative data already collected as part of the enrolment process, and would not necessitate the dissemination and processing of an annual survey among first-year students.

There is much more that can be done in this area to better understand the determinants of these early adverse outcomes at university, and to improve the accuracy of any PRM for identifying students at risk of these poor outcomes. Firstly, we could improve our measures of these early university outcomes. For example, we've concentrated on just the non-completion outcomes for courses. This doesn't distinguish between students who discontinue their study early in the semester (i.e., course dropouts) and those who don't meet the passing standards at the end of the semester (i.e., course failures). Secondly, much more could be done to expand the range of covariates used in the regression analysis. For example, we have no information in our administrative data on parental education, family finances, student scholarships or other financial aid, and peer and community characteristics. We could also do more with existing administrative data to improve the quality of our predictive variables. For example, we have access to only partial information on student academic performance in high school. The national exam results in their last year of high school may be missing for some of the students in our database. It would be possible with available data from the Ministry of Education to gain access to the results from national exams for these students over their two previous years in high school. This could greatly improve the quality of our predictive risk tool, and again help the university in targeting its limited resources at the most vulnerable students.

## References


Altonji, J. G. (1993). "The Demand for and Return to Education when Education Outcomes are Uncertain." *Journal of Labor Economics, 11*(1), 48-83.

Angrist, J., and Lavy, V. (1999). "Using Maimonides's Rule to Estimate the Effect of Class Size on Children's Academic Achievement." *Quarterly Journal of Economics*, 114, 533-575.

Bai, J., and Maloney, T. (2006). "Ethnicity and Academic Success at University." *New Zealand Economic papers, 40*(2), 181-213.

Bean, J. P. (1980). "Dropouts and Turnover. The Synthesis and Test of a Causal Model of Student Attrition." *Research in Higher Education*, 12(2), 155-187.

Belloc, F., Maruotti, A., and Petrella, L. (2010). "University Drop-out: An Italian Experience." *Higher Education*, 60, 127-138.

Betts, J. R., and Morell, D. (1999). "The Determinants of Undergraduate Grade Point Average: The Relative Importance of Family Background, High School Resources, and Peer Group Effects." *Journal of Human Resources, 34*(2), 268-293.

Billings, J., Blunt, I., Steventon, A., Georghiou, T., Lewis, G., and Bardsley, M. (2012). "Development of a Predictive Model to Identify Inpatients at Risk of Re-admission Within 30 Days of Discharge (PARR-30)". *BMJ Open*, 00:e001667.

Billings, J., Dixon, J., Mijanovich, T., and Wennberg, D. (2006). "Case Findings for Patients at Risk of Readmission to Hospital: Development of an Algorithm to Identify High Risk Patients." *BMJ*, doi: 10.1136/bmj.38870.657917.

Cohn, E., Cohn, S., Balch, D. C., and Bradley, J. (2004). "Determinants of Undergraduate GPAs: SAT Scores, High-School GPA and High School Rank." *Economics of Education Review*, 23, 577-586.

Cyrenne, P., and Chan, A. (2012). "High School Grades and University Performance: A Case Study." *Economics of Education Review*, 31, 524-542.

DesJardins, S. L., Ahlburg, D. A., and McCall, B. P. (1999). "An Event History Model of Student Departure." *Economics of Education Review*, 18, 375-390.

DesJardins, S. L., Ahlburg, D. A., and McCall, B. P. (2002). "Simulating the Longitudinal Effects of Changes in Financial Aid on Student Departure from College." *Journal of Human Resources, 37*(3), 653-679.

Dobbelsteen, S., Levin, J., and Oosterbeek, H. (2002). "The Causal Effect of Class Size on Scholastic Achievement: Distinguishing the Pure Class Size Effect from the Effect of Changes in Class Composition." *Oxford Bulletin of Economics and Statistics*, 64, 17-38.

Ficano, C. C. (2012). "Peer Effects in College Academic Outcomes – Gender Matters!" *Economics of Education Review*, 31, 1102-1115.

Grayson, J. P. (1998). "Racial Origin and Student Retention in a Canadian University." *Higher Education*, 36, 323-352.

Grubb, W. N. (1989). "Dropouts, Spells of Time, and Credits in Postsecondary Education: Evidence from Longitudinal Surveys." *Economics of Education Review*, 8, 49-67.

Hartog, J., Pfann, G., and Ridder, G. (1989). "(Non-)Graduation and the Earning Function: An Inquiry on Self-Selection." *European Economic Review*, 33, 1371-1395.

Hoxby, C. M. (2000). "The Effects of Class Size on Student Achievement: New Evidence from Population Variation." *Quarterly Journal of Economics*, 115, 1239-1285.

Ishitani, T. T. (2006). "Studying Attrition and Degree Completion Behaviour Among First-Generation College Students in the United States." *Journal of Higher Education, 77*(5), 861-885.

Kerkvliet, J., and Nowell, C. (2005). "Does One Size Fit All? University Differences in the Influence of Wages, Financial Aid, and Integration on Student Retention." *Economics of Education Review*, 24, 85-95.

Krueger, A. B. (1999). "Experimental Estimates of Education Production Functions." *Quarterly Journal of Economics*, 114(2), 497-532.

Krueger, A. B. (2003). "Economic Considerations and Class Size." *Economic Journal*, 113, F34-F63.

Light, A. (1996). "Hazard Model Estimates of the Decision to Reenrol in School." *Labor Economics*, 2, 381-406.

Light, A., and Strayer, W. (2000). "Determinants of College Completion: School Quality or Student Ability?" *Journal of Human Resources, 35*(2), 299-332.

Manski, C. F. (1989). "Schooling as Experimentation: A Reappraisal of the Postsecondary Dropout Phenomenon." *Economics of Education Review, 8*(4), 305-312.

Mastekaasa, A., and Smeby, J. C. (2008). "Educational Choice and Persistence in Male- and Female-Dominated Fields." *Higher Education*, 55, 189-202.

McPherson, M. S., and Schapiro, M. O. (1991). "Does Student Aid Affect College Enrolment? New Evidence on a Persistence Controversy." *American Economic Review, 81*(1), 309-318.

Montmarquette, C., Mahseredjian, S., and Houle, R. (2001). "The Determinants of University Dropouts: A Bivariate Probability Model with Sample Selection." *Economics of Education Review*, 20, 475-484.

Montmarquette, C., Viennot-Briot, N., and Dagenais, M. (2007). "Dropout, School Performance, and Working While in School." *Review of Economics and Statistics, 89*(4), 752-760.

New Zealand Ministry of Education. (2004). *Retention, Completion and Progression in Tertiary Education 2003.* Wellington: Ministry of Education.

Ost, B. (2010). "The Role of Peers and Grades in Determining Major Persistence in the Sciences." *Economics of Education Review, 29*(6), 923-934.

Rivkin, S. G., Hanushek, E. A., and Kain, J. F. (2005). "Teachers, Schools, and Academic Achievement." *Econometrica, 73*(2), 417-458.

Rodgers, T. (2013). "Should High Non-Completion Rates Amongst Ethnic Minority Students be seen as an Ethnicity Issue? Evidence from a Case Study of a Student Cohort from a British University." *Higher Education, 66*(5), 535-550.

Robst, J., Keil, J., and Russo, D. (1998). "The Effect of Gender Composition of Faculty on Student Retention." *Economics of Education Review, 17*(4), 429-439.

Singell, L. D. (2004). "Come and Stay a While: Does Financial Aid Affect Retention Conditioned on Enrolment at a Large Public University?" *Economics of Education Review*, 23, 459-471.

Stampen, J. O., and Cabrera, A. F. (1988). "The Targeting and Packaging of Student Aid and its Effect on Attrition." *Economics of Education Review, 7*(1), 29-46.

Stinebrickner, T., and Stinebrickner, R. (2012). "Learning About Academic Ability and the Dropout Decision." *Journal of Labor Economics, 30*(4), 707-748.

Stratton, L. S., O'Toole, D. M., and Wetzel, J. N. (2008). "A Multinomial Logit Model of College Stopout and Dropout behaviour." *Economics of Education Review*, 27, 319-331.

Tinto, V. (1982). "Limits of Theory and Practice in Student Attrition." *Journal of Higher Education*, 53, 687-700.

Tinto, V. (1993). *Leaving College: Rethinking the Causes and Cures of Student Attrition*, 2nd Edition. Chicago: Chicago University Press.

Vaithianathan, R., Maloney, T., Putnam-Hornstein, E., and Jiang, N. (2013). "Using Predictive Modelling to Identify Children in the Public Benefit System at High Risk of Substantiated Maltreatment." *American Journal of Preventive Medicine, 45*(3), 354-359.

Venti, S. F., and Wise, D. A. (1982). "Test Scores, Educational Opportunities, and Individual Choice." *Journal of Public Economics*, 18, 35-63.

Wetzel, J. N., O'Toole, D. M., and Peterson, S. (1999). "Factors Affecting Student Retention Probabilities: A Case Study." *Journal of Economics and Finance, 23*(1), 45-55.

**Table 1**
**Descriptive Statistics and Variable Definitions**
**Full Sample**

| Variable | Definition | Mean (std. deviation) |
|---|---|---|
| ***Dependent Variables*** | | |
| **Non-Completion** | 1 if first-year course is completed; zero otherwise | 0.154 (0.361) |
| **Non-Retention** | 1 if student returns to university in the second year; zero otherwise | 0.226 (0.418) |
| ***Year of Cohort*** | | |
| **Year 2009** | 1 if student first enrols in the year 2009; zero otherwise | 0.225 (0.478) |
| **Year 2010** | 1 if student first enrols in the year 2010; zero otherwise | 0.253 (0.435) |
| **Year 2011** | 1 if student first enrols in the year 2011; zero otherwise | 0.240 (0.427) |
| **Year 2012** | 1 if student first enrols in the year 2012; zero otherwise | 0.281 (0.450) |
| ***Ethnicity*** | | |
| **Asian** | 1 if student reports ethnicity as Asian; zero otherwise | 0.242 (0.428) |
| **European** | 1 if student reports ethnicity as European; zero otherwise | 0.392 (0.488) |
| **Māori** | 1 if student reports ethnicity as Māori; zero otherwise | 0.098 (0.297) |
| **Pacifica** | 1 if student reports ethnicity as Pacifica; zero otherwise | 0.112 (0.316) |
| **Others** | 1 if student reports other ethnicity; zero otherwise | 0.080 (0.272) |
| **Unknown** | 1 if students reports no ethnicity; zero otherwise | 0.076 (0.265) |
| ***Country of Origin*** | | |
| **New Zealand** | 1 if student reports New Zealand as country of origin; zero otherwise | 0.695 (0.460) |
| **China** | 1 if student reports China as country of origin; zero otherwise | 0.086 (0.280) |
| **India** | 1 if student reports India as country of origin; zero otherwise | 0.016 (0.127) |
| **Korea** | 1 if student reports Korea as country of origin; zero otherwise | 0.022 (0.147) |
| **Vietnam** | 1 if student reports Vietnam as country of origin; zero otherwise | 0.013 (0.113) |
| **Others** | 1 if student reports other country of origin; zero otherwise | 0.155 (0.362) |
| **Unknown** | 1 if students reports no country of origin; zero otherwise | 0.013 (0.111) |
| ***Personal Characteristics*** | | |
| **Female** | 1 if student is female; zero if male | 0.601 (0.490) |
| **Part-Time** | 1 if student is enrolled part-time; zero full-time | 0.293 (0.455) |
| **Language** | 1 if student reports a first language; zero otherwise | 0.578 (0.494) |
| **English** | 1 if student reports English as first language; zero otherwise (conditional on reporting a first language) | 0.707 (0.455) |
| **Domestic** | 1 if student receives domestic funding; zero otherwise | 0.879 (0.326) |
| **Age** | Mean age | 22.075 (6.322) |
| ***High School Information*** | | |
| **Known NCEA Score** | 1 if NCEA score is available from last year of school; zero otherwise | 0.444 (0.497) |
| **Actual NCEA Score** | Actual NCEA score (conditional on availability of score) | 155.107 (62.860) |
| **Literacy/Numeracy** | 1 if student took literacy and numeracy test in school; zero otherwise | 0.238 (0.426) |
| **School Decile** | Mean school decile (conditional on availability of school decile) | 6.846 (2.812) |

| | | |
|---|---|---|
| ***Entrance Type*** | | |
| **NCEA Admission** | 1 if student entered through NCEA level 3; zero otherwise | 0.363 (0.481) |
| **Special Admission** | 1 if student entered through Special Admission category; zero otherwise | 0.130 (0.336) |
| **Internal** | 1 if student entered through pre-degree at this University; zero otherwise | 0.089 (0.285) |
| **External** | 1 if student entered through study at another university; zero otherwise | 0.150 (0.358) |
| **Cambridge/IB** | 1 if student entered through Cambridge or International Bachelaureate; zero otherwise | 0.014 (0.117) |
| **Others** | 1 if student entered through some other category; zero otherwise | 0.253 (0.435) |

| | | |
|---|---|---|
| ***Course Information*** | | |
| **Study Hours** | Recommended hours of class and preparation time over the semester | 180.539 (62.964) |
| **Known Contact** | 1 if contact hours for the paper are reported | 0.844 (0.362) |
| **Contact Hours** | Contact hours for the paper (conditional on reporting contact hours) | 75.980 (32.234) |
| **Class Size** | Average class size in the course | 38.279(28.932) |
| **Course Size** | Total number of students enrolled in the course | 562.194 (535.464) |
| **Internet Content** | 1 if course is supported with internet content; zero otherwise | 0.588 (0.492) |
| **Level 4** | 1 if course is at level 4 (pre-degree); zero otherwise | 0.005 (0.067) |
| **Level 5** | 1 if course is at level 5 (first year); zero otherwise | 0.836 (0.371) |
| **Level 6** | 1 if course is at level 6 (second year): zero otherwise | 0.156 (0.363) |
| **Level 7** | 1 if course is at level 7 (third year); zero otherwise | 0.004 (0.059) |

| | | |
|---|---|---|
| ***Individual Academic Information*** | | |
| **Number of Courses** | Number of courses taken by the student | 5.243 (2.301) |
| **High Level** | Proportion of level 6 or 7 courses taken by the student | 0.138 (0.202) |
| **Double Degree** | 1 if student is enrolled in a double-degree; zero otherwise | 0.008 (0.088) |
| **One Campus** | 1 if student is taking all courses on a single campus; zero otherwise | 0.913 (0.283) |

| | | |
|---|---|---|
| ***First-Year Programmes of Entry*** | | |
| **BA** | 1 if student enrolled in Bachelor of Arts; zero otherwise | 0.078 (0.268) |
| **BBus** | 1 if student enrolled in Bachelor of Business; zero otherwise | 0.282 (0.450) |
| **BCIS** | 1 if student enrolled in Bachelor of Computer Information Science; zero otherwise | 0.049 (0.216) |
| **BCS** | 1 if student enrolled in Bachelor of Communication Studies; zero otherwise | 0.068 (0.250) |
| **Bde** | 1 if student enrolled in Bachelor of Design | 0.074 (0.262) |
| **BEdu** | 1 if student enrolled in Bachelor of Education | 0.040 (0.197) |
| **BEngT** | 1 if student enrolled in Bachelor of Engineering Technology; zero otherwise | 0.029 (0.168) |
| **BHS** | 1 if student enrolled in Bachelor of Health Science; zero otherwise | 0.195 (0.396) |
| **BIHM** | 1 if student enrolled in Bachelor of International Hospitality Management; zero otherwise | 0.043 (0.204) |
| **BMS** | 1 if student enrolled in Bachelor of Mathematical Science; zero otherwise | 0.006 (0.079) |
| **BSR** | 1 if student enrolled in Bachelor of Sports and Recreation; zero otherwise | 0.060 (0.237) |
| **Others** | 1 if student enrolled in another smaller programme; zero otherwise | 0.075 (0.291) |

**Table 2**
**Estimated Results from Maximum Likelihood Probit Analysis on Course Non-Completion and Student Non-Retention Estimation Subsample**

| | Course Non-Completion in First Year | | | Student Non-Retention in Second Year | | |
|---|---|---|---|---|---|---|
| **Variable** | **Coefficient** | **Std. Error** | **dy/dx** | **Coefficient** | **Std. Error** | **dy/dx** |
| **Constant** | -0.6693*** | 0.1148 | - | -0.1374 | 0.2156 | - |
| ***Year of Cohort*** | | | | | | |
| **Year 2009** | -0.1069*** | 0.0217 | -2.20% | -0.0134 | 0.0447 | -0.36% |
| **Year 2010** | -0.0752*** | 0.0198 | -1.54% | 0.1030** | 0.0424 | 2.74% |
| **Year 2011** | -0.1725*** | 0.0202 | -3.54% | 0.1112*** | 0.0428 | 2.96% |
| ***Ethnicity*** | | | | | | |
| **Asian** | -0.1267*** | 0.0442 | -2.60% | -0.1079 | 0.0901 | -2.87% |
| **European** | -0.1830*** | 0.0483 | -3.76% | 0.0164 | 0.0984 | 0.44% |
| **Māori** | 0.1666*** | 0.0515 | 3.42% | 0.2200** | 0.1064 | 5.85% |
| **Pacifica** | 0.3247*** | 0.0501 | 6.67% | 0.0967 | 0.1040 | 2.58% |
| **Others** | -0.0417 | 0.0511 | -0.86% | -0.0525 | 0.1048 | -1.40% |
| ***Country of Origin*** | | | | | | |
| **New Zealand** | -0.4828*** | 0.0630 | -9.91% | -0.7431*** | 0.1258 | -19.78% |
| **China** | -0.4488*** | 0.0714 | -9.21% | -0.8090*** | 0.1411 | -21.53% |
| **India** | -0.4670*** | 0.0865 | -9.59% | -0.7482*** | 0.1777 | -19.91% |
| **Korea** | -0.3356*** | 0.0809 | -6.89% | -0.4329*** | 0.1616 | -11.52% |
| **Vietnam** | -0.6662*** | 0.1081 | -13.68% | -1.5075*** | 0.2574 | -40.12% |
| **Others** | -0.4833*** | 0.0650 | -9.92% | -0.8868*** | 0.1299 | -23.60% |
| ***Personal Characteristics*** | | | | | | |
| **Female** | -0.1328*** | 0.0165 | -2.73% | -0.0583* | 0.0343 | -1.55% |
| **Part-Time** | 0.7546*** | 0.0179 | 15.49% | 0.5935*** | 0.0466 | 15.80% |
| **Language** | 0.0302 | 0.0250 | 0.62% | 0.0304 | 0.0523 | 0.81% |
| **English** | 0.0057 | 0.0269 | 0.12% | -0.0144 | 0.0568 | -0.38% |
| **Domestic** | 0.1852*** | 0.0434 | 3.80% | 0.2495*** | 0.0880 | 6.64% |
| **Under Age 18** | 0.1244 | 0.1098 | 2.55% | -0.2035 | 0.2180 | -5.42% |
| **Age 18** | 0.1854*** | 0.0645 | 3.81% | -0.1700 | 0.1243 | -4.53% |
| **Age 19** | 0.2233*** | 0.0646 | 4.58% | -0.0735 | 0.1245 | -1.96% |
| **Age 20** | 0.3473*** | 0.0649 | 7.13% | -0.0037 | 0.1250 | -0.10% |
| **Age 21** | 0.3418*** | 0.0658 | 7.02% | 0.0469 | 0.1272 | 1.25% |
| **Age 22** | 0.2411*** | 0.0679 | 4.95% | -0.0147 | 0.1323 | -0.39% |
| **Age 23** | 0.2555*** | 0.0695 | 5.25% | -0.082 | 0.1353 | -2.18% |

| | | | | | | |
|---|---|---|---|---|---|---|
| **Age 24** | 0.1512** | 0.0730 | 3.10% | -0.0653 | 0.1421 | -1.74% |
| **Age 25** | 0.1427* | 0.0758 | 2.93% | 0.0822 | 0.1439 | 2.19% |
| **Ages 26 to 30** | 0.0764 | 0.0665 | 1.57% | 0.0286 | 0.1251 | 0.76% |
| **Ages 31 to 35** | 0.0264 | 0.0730 | 0.54% | -0.1324 | 0.1364 | -3.52% |
| **Ages 36 to 45** | 0.0838 | 0.0720 | 1.72% | -0.1190 | 0.1365 | -3.17% |
| ***High School Information*** | | | | | | |
| **Known NCEA Score** | 0.3703*** | 0.0344 | 7.60% | 0.1991*** | 0.0741 | 5.30% |
| **Actual NCEA Score** | -0.0047*** | 0.0002 | -0.10% | -0.0024*** | 0.0005 | -0.06% |
| **Literacy/Numeracy** | 0.0819*** | 0.0259 | 1.68% | 0.1404*** | 0.0473 | 3.74% |
| **School Decile 1** | 0.4710*** | 0.0423 | 9.67% | 0.084 | 0.0938 | 2.23% |
| **School Decile 2** | 0.2370*** | 0.0419 | 4.87% | -0.1128 | 0.0899 | -3.00% |
| **School Decile 3** | 0.1823*** | 0.0374 | 3.74% | 0.0225 | 0.0804 | 0.60% |
| **School Decile 4** | 0.1674*** | 0.0326 | 3.44% | -0.1706** | 0.0696 | -4.54% |
| **School Decile 5** | 0.0865** | 0.0391 | 1.78% | -0.0602 | 0.0812 | -1.60% |
| **School Decile 6** | 0.0740** | 0.0374 | 1.52% | -0.1361* | 0.0778 | -3.62% |
| **School Decile 7** | 0.0886*** | 0.0340 | 1.82% | -0.1345* | 0.0718 | -3.58% |
| **School Decile 8** | 0.1264*** | 0.0349 | 2.60% | 0.0019 | 0.0724 | 0.05% |
| **School Decile 9** | 0.1415*** | 0.0323 | 2.91% | -0.063 | 0.0664 | -1.68% |
| **School Decile 10** | 0.1607*** | 0.0289 | 3.30% | -0.1293** | 0.0601 | -3.44% |
| ***Entrance Type*** | | | | | | |
| **NCEA Admission** | 0.1752*** | 0.0363 | 3.60% | -0.0128 | 0.0768 | -0.34% |
| **Special Admission** | 0.0752*** | 0.0270 | 1.54% | -0.0827 | 0.0559 | -2.20% |
| **Internal** | -0.0312 | 0.0310 | -0.64% | -0.2004*** | 0.0653 | -5.33% |
| **External** | -0.1542*** | 0.0274 | -3.17% | -0.1752*** | 0.0556 | -4.66% |
| **Cambridge/IB** | -0.3138*** | 0.0703 | -6.44% | -0.3563** | 0.1517 | -9.48% |
| ***Course Information*** | | | | | | |
| **Study Hours** | 0.0030 | 0.0265 | 0.06% | - | - | - |
| **Known Contact** | -0.0702** | 0.0346 | -1.44% | - | - | - |
| **Contact Hours** | -0.0101 | 0.0480 | -0.21% | - | - | - |
| **Class Size/10** | 0.0071** | 0.0028 | 0.10% | - | - | - |
| **Course Size/10** | -0.0015*** | 0.0003 | -0.03% | - | - | - |
| **Internet Content** | 0.0187 | 0.0187 | 0.38% | - | - | - |
| **Level 4** | -0.2765*** | 0.1049 | -5.68% | - | - | - |
| **Level 6** | -0.0218 | 0.0228 | -0.45% | - | - | - |
| **Level 7** | -0.1097 | 0.1164 | -2.25% | - | - | - |
| ***Individual Academic Information*** | | | | | | |
| **Number of Courses** | - | - | - | -0.0010 | 0.0119 | -0.03% |
| **6+ Courses** | - | - | - | -0.7818*** | 0.0963 | -20.81% |
| **Double Degree** | -0.1524 | 0.0936 | -3.13% | -0.4927** | 0.2423 | -13.11% |

| | | | | | | |
|---|---|---|---|---|---|---|
| **One Campus** | -0.0075 | 0.0247 | -0.15% | 0.0446 | 0.0559 | 1.19% |
| | | | | | | |
| ***First-Year Programmes of Entry*** | | | | | | |
| **BA** | 0.0837*** | 0.0316 | 1.72% | 0.3280*** | 0.0748 | 8.73% |
| **BBus** | -0.1791*** | 0.0469 | -3.68% | -0.1242** | 0.0719 | -3.30% |
| **BCIS** | 0.0167 | 0.0372 | 0.34% | -0.0737 | 0.0868 | -1.96% |
| **BCS** | -0.2971*** | 0.0410 | -6.10% | -0.1786** | 0.0887 | -4.75% |
| **BDe** | -0.3306*** | 0.0397 | -6.79% | -0.1221 | 0.0818 | -3.25% |
| **BEdu** | -0.5306*** | 0.0456 | -10.89% | -0.2231** | 0.0948 | -5.94% |
| **BEngT** | 0.0896* | 0.0462 | 1.84% | -0.1237 | 0.1064 | -3.29% |
| **BHS** | -0.3598*** | 0.0328 | -7.39% | -0.033 | 0.0642 | -0.88% |
| **BIHM** | -0.2196*** | 0.0407 | -4.51% | -0.0913 | 0.0930 | -2.43% |
| **BMS** | 0.1758** | 0.0727 | 3.61% | 0.1497 | 0.1930 | 3.98% |
| **BSR** | -0.1145*** | 0.0366 | -2.35% | 0.1480* | 0.0795 | 3.94% |

| | | |
|---|---|---|
| **Pseudo $R^2$** | 0.1339 | 0.1063 |
| **Log-Likelihood** | -18,985.6 | -4,417.61 |
| **Area Under the ROC Curve** | 0.7553 | 0.7125 |
| ***n*** | 50,932 | 9,301 |

Notes: *** Indicates significance at the 1% level
** Indicates significance at the 5% level
* Indicates significance at the 10% level

**Figure 1**
**Area Under the ROC Curve**
**Course Non-Completion in the First Year**

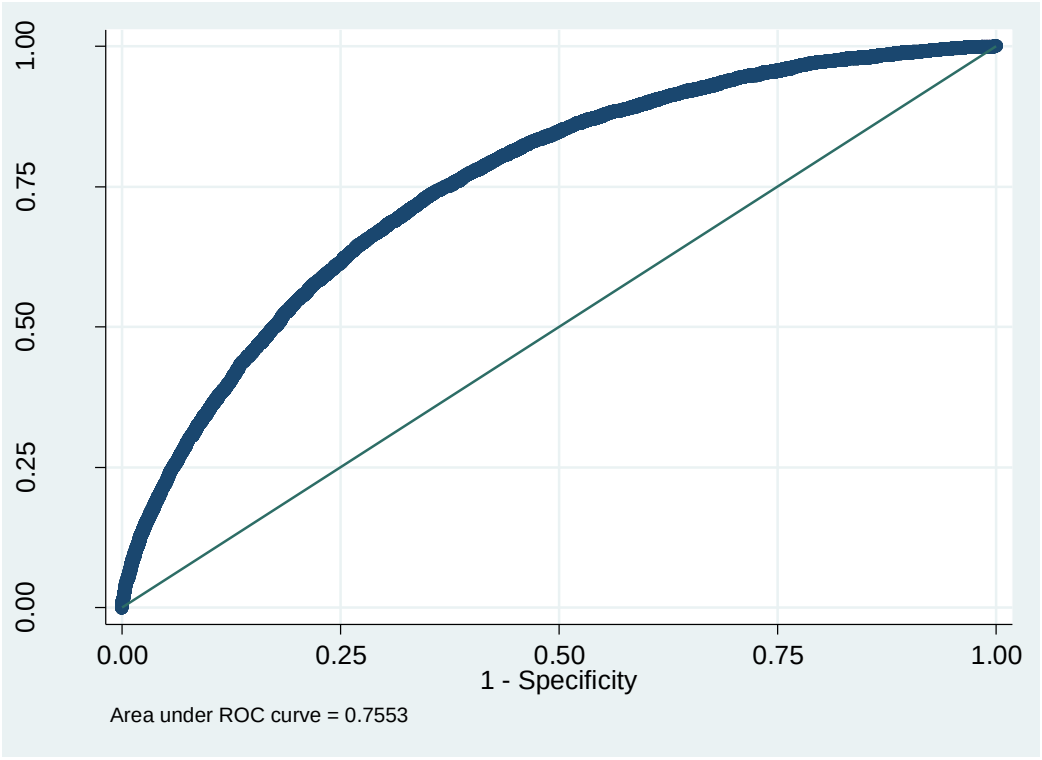
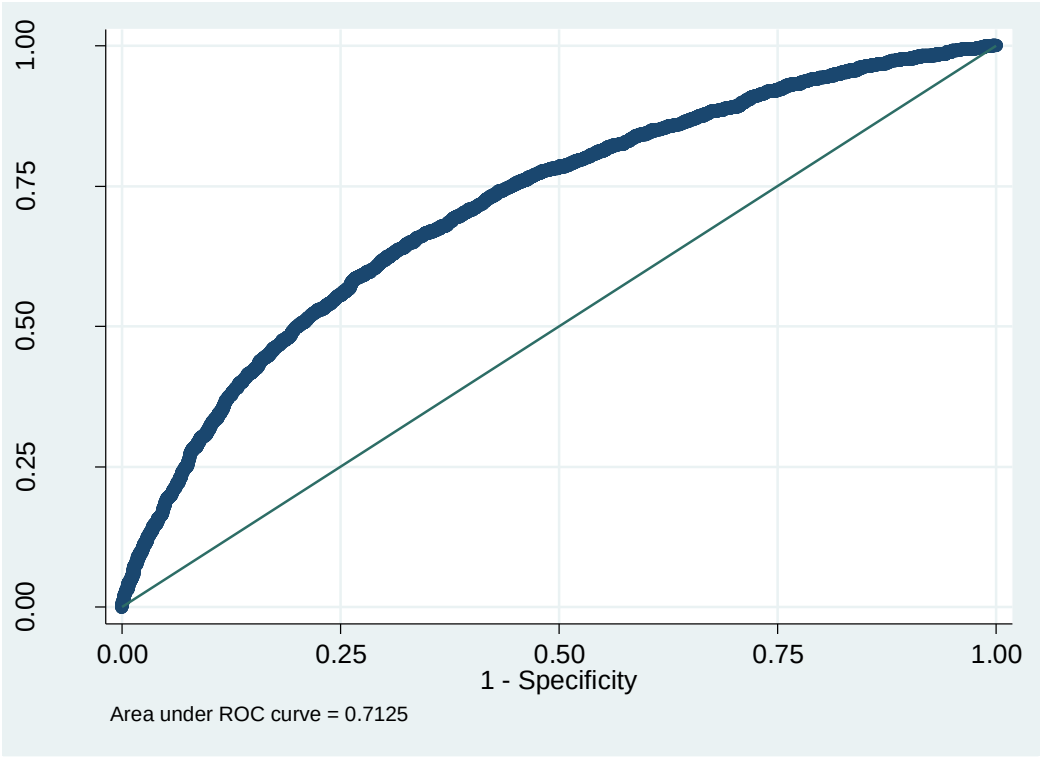



**Figure 2**
**Area Under the ROC Curve**
**Student Non-Retention in the Second Year**



| Table 3<br>Percentage of Outcomes Correctly Identified<br>Validation Subsample | | |
|---|---|---|
| | **Course Non-Completion in First Year** | **Student Non-Retention in Second Year** |
| **Top 1 Decile (top 10%)** | 29.25% | 23.33% |
| **Top 2 Deciles (top 20%)** | 47.57% | 40.91% |
| **Area Under the ROC Curve** | 0.7553 | 0.7125 |
| ***n*** | 50,932 | 9,301 |