

---

## Enrolment Management in Graduate Business Programs: Predicting Student Retention

ABDOLREZA ESHGHI,<sup>a</sup> DOMINIQUE HAUGHTON,<sup>a,b</sup> MINGFEI LI,<sup>a</sup> LINDA SENNE,<sup>a</sup>

MARIA SKALETSKY<sup>a</sup> AND SAM WOOLFORD<sup>a</sup>

<sup>a</sup> *DART (Data Analytics Research Team), Bentley University, USA*

<sup>b</sup> *Université Toulouse 1 Capitole, France*

Submitted to the *Journal of Institutional Research* August 11, 2011, accepted for publication October 12, 2011

---

---

### Abstract

The increasing competition for graduate students among business schools has resulted in a greater emphasis on graduate business student retention. In an effort to address this issue, the current article uses survival analysis, decision trees and TreeNet® to identify factors that can be used to identify students who are at risk of dropping out of a graduate business program. This work extends the literature in several ways. First, it looks at attrition among business school graduate students. Second, because graduate business education typically involves a mix of full-time and part-time students, our study incorporates both these groups. Finally, we use methodologies (survival analysis with time-dependent predictors, decision trees and TreeNet®) which, to the best of our knowledge, have not been employed previously for studying student retention. Our results uncover several factors that could help administrators develop intervention strategies to increase graduate business student retention.

**Keywords:** Graduate student retention, graduate business programs, survival analysis, decisions trees, Treenet®

---

### Background

The graduate business education marketplace has undergone significant changes in recent years. First, the spiralling cost of attending private and public universities has made graduate business education unreachable for many. Increased costs coupled with dwindling corporate sponsorships, scholarships, and financial assistance at federal and state levels have significantly limited the pool of prospective students for Master in Business Administration (MBA) and Master of Science (MS) programs in business. Consequently, more students are enrolling in part-time and evening programs rather than pursuing full-time studies. Second, many employers are increasingly opting for custom-designed corporate MBAs, which further cut into the pool of applicants to business schools (DeShields et al., 2005). Third, the information technology revolution has led to a proliferation of online courses and digital

universities, providing more choices to those who seek a graduate degree in business (Friga et al., 2003).

The combination of these factors has led to intense competition in the market for graduate business education. Faced with these realities, colleges and universities have responded by adopting the principles of market orientation, whereby more emphasis is placed on understanding the needs of the students in an attempt to create superior educational value for them (Hammond et al., 2006). Moreover, the typical enrolment function with a focus on securing a healthy enrolment level is being replaced by a senior-level position charged with the responsibility of focusing not only on securing healthy enrolment levels, but also on paying attention to students' experiences while they are enrolled in a degree program. In other words, colleges and universities are increasingly focusing on all phases of the student lifecycle from prospects to students to alumni, rather than just identifying prospects and recruiting students. As a result, retention of existing students has become a top priority.

In the for-profit sector, where firms face similar competitive circumstances, executives have created strategies for managing the customer lifecycle—managing the distinct stages that customers go through from the day they become prospects to the day they cease doing business with a company in an effort to optimise both customer acquisition and customer retention (Rust, Zeithaml, & Lemone, 2000). We propose that the concept of customer lifecycle is relevant for enrolment management in graduate schools of business and that optimising student retention can have significant marketing and financial implications since it costs more to recruit and admit new students than to retain current students.

Therefore, the overall focus of this article is on student retention and, in particular, to the identification of students at risk. If at-risk students—students who are likely to drop out of their graduate programs—can be identified then it may be possible for business schools to develop intervention strategies to improve student retention.

### **Literature**

Several authors (Astin, 1997; Braunstein et al., 2006; DeShields et al., 2005; Druzdzel & Glymour, 1994; Johnes & McNabb, 2004; Marcus, 1989; Sanders & Burton, 1996) have explored undergraduate student attrition. The majority of these studies have considered characteristics associated with undergraduate attrition using some form of regression-related analysis (either linear, logistic, two-stage least squares or path analysis) typically using a cohort analysis. None of these studies explicitly considers time as a factor in their analysis.

Other authors including Ott et al. (1984) and Stock et al. (2006) have considered graduate student attrition at the masters or doctoral level, also using regression-based analysis with cohorts. Booth and Satchell (1995) use a competing risks model to evaluate the effect of public funding on the retention of British doctoral students.

None of the studies that we found in the literature focused on the attrition of business school students and, in particular, none attempted to consider part-time students. In addition, none of the studies attempted to include time explicitly as a factor in modelling attrition.

This article expands on past work in several ways. First, it investigates attrition among business school graduate students. Second, because graduate business education typically involves a mix of full-time and part-time students, both full-time and part-time students are included in this study. Including part-time students introduces an additional

complication in any model relating student characteristics to retention in that it may not be possible to determine if a student has dropped out or is a part-time student who has temporarily interrupted his/her studies but will, in fact, come back to complete his/her degree. For that reason, we have defined the ‘dropout’ status and the length of time until dropout in a distinctive way; this adds to existing literature where this difficulty has not been addressed so far to the best of our knowledge.

The next section describes the methodologies we used, including a detailed description of the dataset and how we define whether a student is still active (has not dropped out) at any point in time. We then present the results, followed by summary and discussion where we examine the implications of our results. The more technical details of our analyses are presented in an Appendix.

### **Dataset**

Data we used in this study was collected from 2,275 students enrolled from January 2001 to August 2007 at a private business school in the northeast of the United States, henceforth referred to as ‘the university’. The students were enrolled for at least one course in the university’s Masters of Business Administration and business-related master’s programs (such as Finance, Marketing, etc.) during the period from January 2001 through August 2007. The duration of an MBA program is typically of about two academic years (the equivalent of four semesters, or four terms) assuming full-time attendance, corresponding to 18 three-credit courses. For example, an MBA study plan in the northern hemisphere could include four semesters (terms) with four courses each and a summer term with two additional courses. Students with prior business academic qualifications can often complete an MBA program more rapidly, and students who pursue their MBA degree part-time will typically study for three or more academic years. The typical duration of a business MS program is three semesters (terms), inclusive of about 10 three-credit courses.

The dataset includes four types of data: administrative (degree program, full-time or part-time status, etc.), demographic (gender, marital status, ethnicity, etc.), academic background (GMAT score, prior GPA, etc.), and academic performance data (course enrolments, graduate program grades etc.). It is important to note that since many students pursue their degree on a part-time basis, student cohorts cannot be defined for this data. To the best of our knowledge, this article contains the first attempt to analyse retention when cohorts cannot be defined. This is important, since many business universities enrol a mix of full-time and part-time students.

In this analysis, we consider the period from January 2001 (beginning of spring term) through August 2007 (end of summer term) and we measure time in units of school terms, yielding a total of 20 school terms. In addition to traditional fall and spring terms, the university has two relatively short summer terms, one that begins in May and a second that begins in July. For analysis purposes, we merged these two sessions into one summer term. The period of twenty school terms as defined above is long enough to identify students who have dropped out even if they are pursuing their degree part-time. To differentiate the students in the dataset, we identified three possible groups:

- **Group 1—Graduated Students:** The students in this group completed a degree before the end of term 20.
- **Group 2—Active Students:** This group consisted of students who had not graduated before term 20 but did register at least once in terms 17 through 20. Since it is

unknown if this group of students will complete their studies, the dropout status for this group is censored.

Note that this definition does not require a student to be continuously registered. A student could have disappeared between terms 5 and 18, for example, but reappeared in term 18—such a student is defined as active.

The length of time active students are defined to be actively pursuing their degree is calculated as:

$$20 - \text{first term-registered in dataset} + 1$$

where 20 is the number of terms in the dataset.

- **Group 3—Inactive Students.** These students did not register once in the last four term blocks in the dataset. We assume that they stopped pursuing their degrees. For inactive students, the length of time until they drop out is:

$$\text{Last term registered} - \text{first term registered} + 1.$$

The length of time defined for groups 2 and 3 is used as a dependent (target) variable in the survival analysis. In group 2 we know that the student has ‘survived’ (has not dropped out) for at least that length of time, but we do not know how much longer the student will ‘survive’. This is why students in group 2 are considered to be ‘censored’.

Since our analysis is focused on the characteristics of students who drop out, we exclude all the students who graduated from our analysis dataset. For all remaining students we include a variable indicating whether they are active or inactive and another capturing the number of terms they have been pursuing a degree in the study period.

Our intent is to determine if readily available data from a student’s application or course performance could help predict the risk of a student dropping out. Consequently, we limit ourselves to variables typically available from the registrar in order to identify covariates associated with the risk of dropping out. After performing exploratory analysis on these variables to determine those covariates that are appropriate for further analysis (e.g., do not involve large numbers of missing values, have a reasonable distribution of values), we included the following covariates in our analysis:

- degree program (MBA or MS)
- status (full-time or part-time)
- marital status (married or not married)
- citizenship (US or not)
- visa (visa holder or not)
- ethnicity (Caucasian, Asian, black/Hispanic/Native American, other/unknown)
- gender (male or female)
- GPA (grade point average at application)
- age (at acceptance to the degree program)
- cumulative GPA (cumulative grade point at the university)

The dataset provided by the Registrar included 5,030 students. Removing the MBA and MS candidates who had graduated yielded a group of 3,435 students. After we eliminated students who switched from part-time to full-time status during the study period from this subset, there were 2,275 students left. These students make up the dataset we analysed to identify at-risk students. The number of semesters in school for students in our dataset who

dropped out is about 5 on average, with a right-skewed distribution: about three fourths of students who have dropped out tend to have done so after, at most, 7 semesters in school. We note that this may be a slight underestimation, since a student who is reported to have attended term 1 (spring term 2001) in our dataset may in fact have been in school for an unknown number of terms before term 1.

### Analytical Approach

To identify characteristics associated with at-risk graduate students we used three different methodologies to examine data on 2,275 students collected from January 2001 to August 2007. The three methodologies we applied are described below.

#### Survival Analysis

In considering why students drop out of a graduate program, survival analysis allows us to model not only whether a student drops out, but also when the event occurs by estimating the risk of a student dropping out at any particular time. This kind of analysis represents an enhancement over linear regression or logistic regression analyses by explicitly including time in the analysis. Survival analysis has been used in other industries to track customer attrition (see for instance Lu, 2002).

We applied a Cox proportional hazards model to the data (Cleves et al., 2004; Hosmer & Lemeshow, 1999). The general form of the model is given as

$$h(t|X) = h_0(t) \exp[\beta_1 X_1 + \beta_2 X_2 + \dots + \beta_p X_p] = h_0(t) r(X, \beta)$$

where  $h(t|X)$  is the hazard function at time  $t$  given the covariates  $X$  and represents the probability that a student who is active up to time  $t$  will drop out at time  $t$  given the covariate values  $X = (X_1, X_2, \dots, X_p)$ . Here  $\exp(\cdot)$  is the exponential function, and  $h_0(t)$  is the baseline hazard at time  $t$  corresponding to  $X = 0$ . One can also define a survival function corresponding to the hazard function as

$$S(t|X) = \exp[-\int_0^t h(s|X) ds]$$

where  $S(t|X)$  is the probability that a student does not drop out until after time  $t$  (i.e., survives beyond time  $t$ ) given the covariate values  $X = (X_1, X_2, \dots, X_p)$ . We also utilise an extended Cox model in order to include the dynamic factor—cumulative GPA—which changes from term to term for each student.

#### Decision Trees

Decision Trees attempt to partition the space represented by a set of ‘predictor variables’ to better discriminate among the values of a target variable. The resulting partitions, representing decision rules, represent combinations of predictor variable values that are associated with specific values of the target variable. We utilised two decision tree methods: Chi Square Automatic Interaction Detector (CHAID) and Classification and Regression Trees (C&RT). CHAID was originally introduced for use with categorical predictor variables while C&RT is suitable for use with continuous as well as categorical predictor variables.

#### TreeNet

TreeNet is a tool typically used after a dataset has been explored with tools like C&RT that enable analysts to refine their models further. According to Salford Systems

(Salford Systems, 2010), in most cases TreeNet will confirm the primary findings reported by C&RT while substantially increasing the predictive accuracy of the models. In our case, the TreeNet analysis provides a more precise understanding of the non-linear relationships between student characteristics and the propensity to drop out of graduate school and helps identify interactions between the predictors of dropping out.

In utilising the above methodologies, we attempted to identify the characteristics most associated with students who drop out. We then compared our results across analyses to identify a consensus around factors that increase the risk of a student dropping out. We present our main findings in the next sections. Technical details of the analyses are presented in the Appendix.

## Findings

The exploratory survival analysis identified the characteristics of at-risk students as summarised in Table 1. The survival analysis results indicate a high degree of consistency between the Cox model and the extended Cox model, indicating stability of the models. The Cox model parameter estimates are provided in Table 2.

**Table 1**

*Likelihood of Dropping out of Graduate School: Univariate Analyses*

Variable	Probability of Dropping Out	
	Higher Risk	Lower Risk
Degree program	MS	MBA
Status	Full-time	Part-time
Marital status	Married	Not married
Age	Older (risk ↑ as age ↑)	Younger (risk ↓ as age ↓)
GPA	Lower GPA (risk ↑ as GPA ↓)	Higher GPA (risk ↓ as GPA ↑)
Citizenship	Non-citizens	Citizens
Visa	Visa students	Non-visa students
Ethnicity	Other ethnic groups	Caucasians

**Table 2**

*Cox Model Parameter Estimates*

Variables in the Equation	Regression Coefficient ( <i>b</i> )	Stand. Err. <i>SE</i> ( <i>b</i> )	Sig.
Degree program (1 if MS, 0 if MBA)	.248	.049	.000
Status (1 if part-time, 0 if full-time)	-1.197	.074	.000
Marital status (1 if married, 0 if not)	.214	.096	.026
Age	.036	.005	.000
GPA	-.390	.056	.000
Degree X Age	-.038	.006	.000

*Note.* *b* is the regression coefficient; *SE*(*b*) is the standard error of the regression coefficient

The following conclusions can be drawn from the results in Table 2:

- **Full-time/part-time status**  
The parameter associated with full-time/part-time status of -1.197 indicates that full-time students have a higher risk of dropping out than part-time students assuming all other covariates are fixed. The risk of dropping out is .302 times lower for a part-time student after controlling for all other covariates. The result may reflect the fact that part-time students have greater flexibility in managing their course load and so find it easier to maintain their status in the program.
- **Marital status**  
The coefficient for marital status of .214 indicates that the risk of dropping out is 1.21 times higher for married students than for unmarried students, possibly reflecting the fact that married students have additional family obligations on top of an academic program.
- **Entering GPA**  
The coefficient associated with entering GPA indicates that a student entering with a one unit higher GPA score has a .61 times lower risk of dropping out. Perhaps better prepared students have an easier time pursuing a graduate degree.
- **Age and degree**  
In order to interpret the impact of age and degree, we must consider the interaction as well as the main effects. For students pursuing an MS, age has very little impact on the risk of dropping out; this risk is .998 times lower for each year older a student is at entry into the program. MBA students, on the other hand, show a slightly higher risk (1.04 times) of dropping out for each year increase in entering age, all other covariates fixed.

Besides the ‘snap shot’ factors that we considered in the previous Cox proportional hazards model (degree program, status, marital status, age and GPA at enrolment), a specific dynamic factor—cumulative GPA—for each term was also studied using an extended Cox model. The resulting parameter estimates are provided in Table 3 below:

**Table 3**

*Extended Cox Model for Cumulative GPA*

Variables in the Equation	Regression Coefficient (b)	Stand. Err. SE(b)	Sig.
Degree program (1 if MS, 0 if MBA)	0.339	0.049	.000
Status (1 if part-time, 0 if full-time)	-0.902	0.073	.000
Marital status (1 if married, 0 if not)	0.184	0.096	.056
Age	0.058	0.005	.000
GPA	-0.248	0.058	.000
Degree X Age	-0.049	0.006	.000
Cumulative GPA	-0.555	0.019	.000

The result of this extended Cox model confirms the same significant impact of full-time/part-time status, degree program, GPA at registration, the age of the student and the interaction of the degree and age on the students’ risk of dropping out. Marital status has a marginal impact ( $p = .056$ ) on the risk of dropping out with all other covariates held constant.

In addition, the student’s cumulative GPA in school has a significant impact on the risk of dropping out. The parameter associated with cumulative GPA is  $-.555$ , indicating that, in a given term, students who have lower cumulative GPAs have a higher risk of dropping out than those who have higher cumulative GPAs, assuming that all other covariates are fixed. After controlling for all other covariates, the risk of dropping out is 0.574 times lower for a student who has one unit higher cumulative GPA at the university in a given term. Perhaps a higher cumulative GPA in school makes students more confident that they will be successful in continuing to pursue their degree in the program. While it may seem intuitive that ‘better’ students are less likely to drop out, we have not found it explicitly demonstrated in the literature.

Both the Cox model and extended Cox model reveal a more complicated relationship between age and degree program and both suggest that the risk of attrition shows a slight increase with age for MBA students. For MS students, the risk of attrition goes up with age in the extended Cox model and is also roughly independent of age. In both models, the risk of attrition for MS students is higher than that for MBA students for the same age student. The extended Cox model also indicates that risk of attrition increases with decreasing cumulative GPA.

As stated earlier, we used decision trees to identify classification rules that best differentiate students who drop out from those that do not. After creating initial trees and pruning as appropriate, both the CHAID and C&RT trees found age, degree program and status to be the best predictors of whether a student was active or not. In addition, both trees concur: the propensity to drop out increases with age overall, with those in their mid-thirties particularly at risk (see for example Figure 1, about 91% dropout rate in Node 5). Interestingly CHAID identifies a decrease in the dropout rate for age beyond 36 or so (about 86% in Node 6); this was confirmed in the TreeNet® analysis described later in the article. Note that the average age in the dataset is close to 35, so that Node 6—with centred ages of 1.15 or greater—involves students who are about 36 or more. The trees also reveal, as was found in the survival analysis, that the propensity for dropping out is higher for MS degrees than MBA degrees for the youngest students. The trees clearly indicate an interaction between age and degree. Note an at-risk group identified by C&RT is younger (less than 32 or so) full-time MS candidates (100% drop out rate, Node 8, Figure 2).

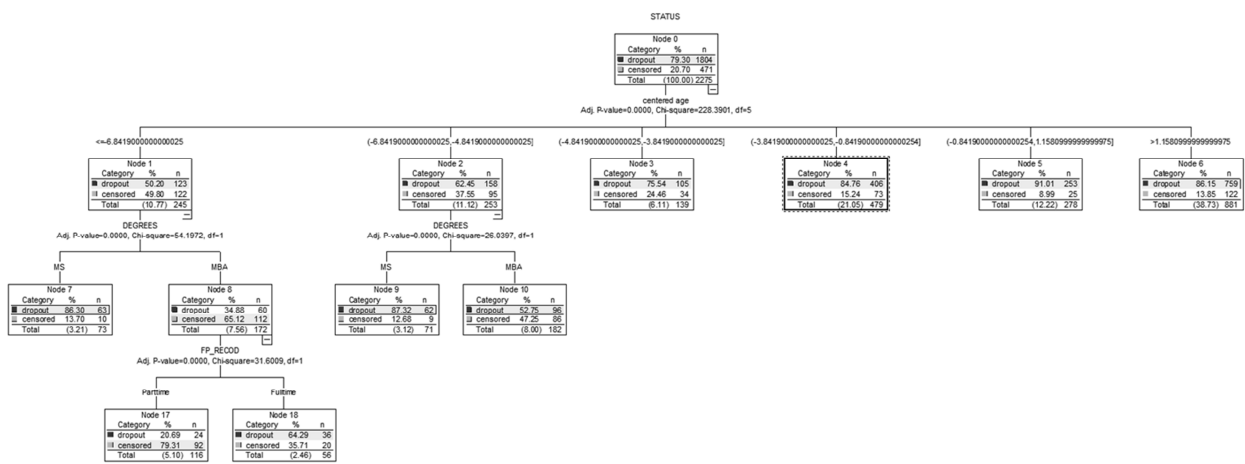


Figure 1. CHAID tree.



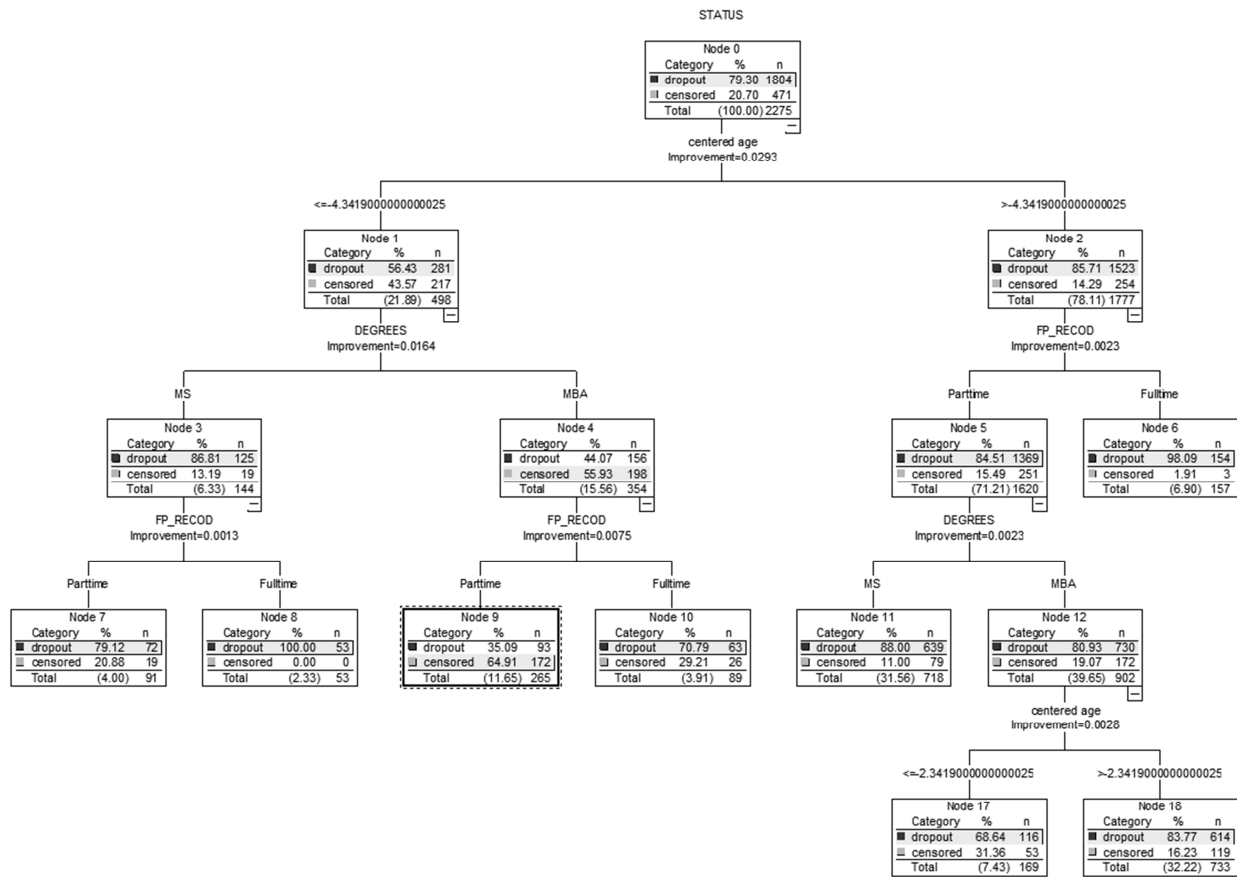


Figure 2. C&RT tree.

Finally, we outline the results of a TreeNet<sup>®</sup> (TreeBoost, see Friedman, 1999) analysis where the target (dependent) variable is whether a student is active or not and the predictors are chosen to be age, GPA, status, marital status and degree program. The predictors were chosen to coincide with predictors that were identified as important in the earlier survival analysis (see Table 4 below)

Table 4

Variable Importance in the TreeNet<sup>®</sup> Model

Variable	Importance
Age	100.00
GPA	85.34
Degree Program	49.79
Part-time/Full-time Status	44.92
Marital status	30.29

Using the TreeNet<sup>®</sup>, we obtained a more precise understanding of the relationship between the target variables and the predictors. As shown in Figure 3, propensity to drop out sharply increases for ages up to near early 40s, is approximately constant, and then decreases slightly for more advanced ages (such as 50 and above). But, in Figure 4, we can see that the propensity to drop out essentially vanishes for GPAs of about three and above.

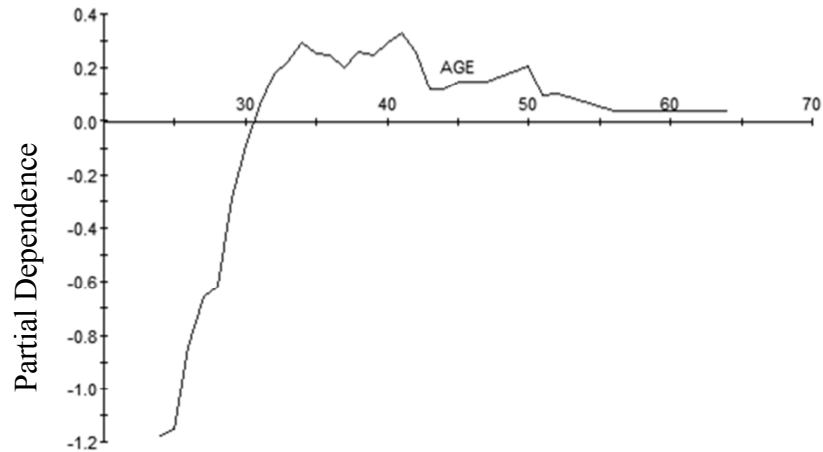


Figure 3. Partial effect of age on the dropout rate (controlling for other predictors).

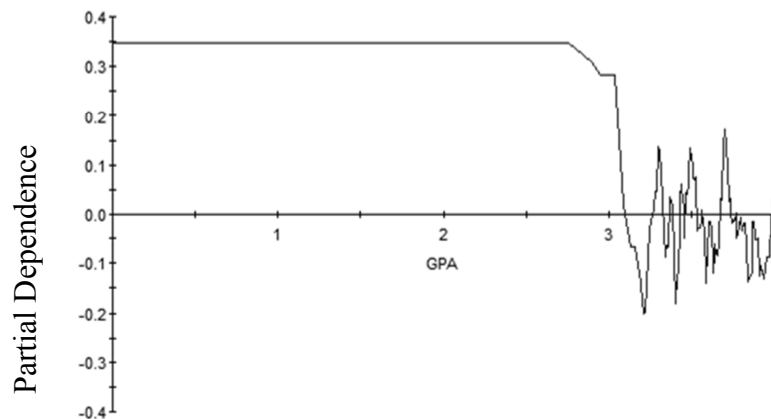


Figure 4. Partial effect of GPA on the dropout rate (controlling for other predictors).

Figure 5 reveals that the propensity to drop out is higher for MSs (coded 2) than for MBAs (coded 1), higher for full-time students (coded 0) than for part-time students (coded 1), and higher for married students (coded 1), possibly because of conflicting family demands.

The next four graphs in Figure 6 display interaction effects among pairs of variables. As we can see, the effect of GPA on the propensity to drop out does vary with age and the difference between MS students and MBA students is particularly sharp for full-time students and much less so for part-time students. We can also see that the differences in propensities to drop out as a function of marital status are more noticeable for MS students than for MBA students. Finally, we note that the propensity to drop out is lower for part-time than full-time students independently of marital status.

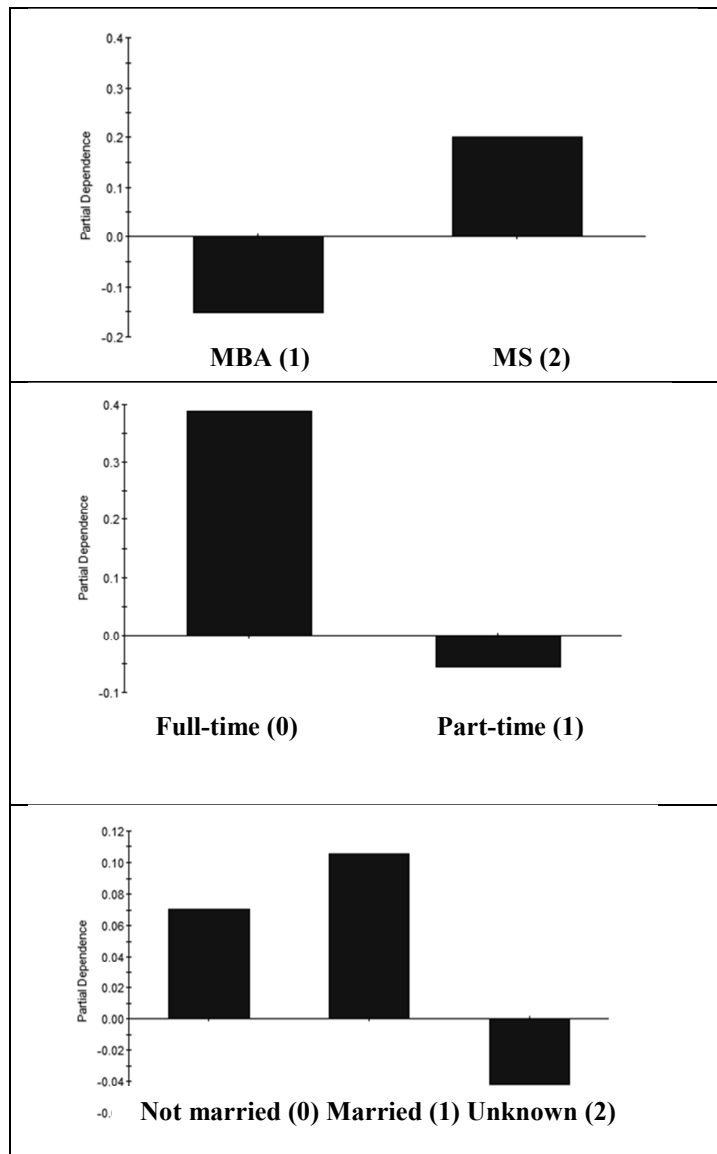
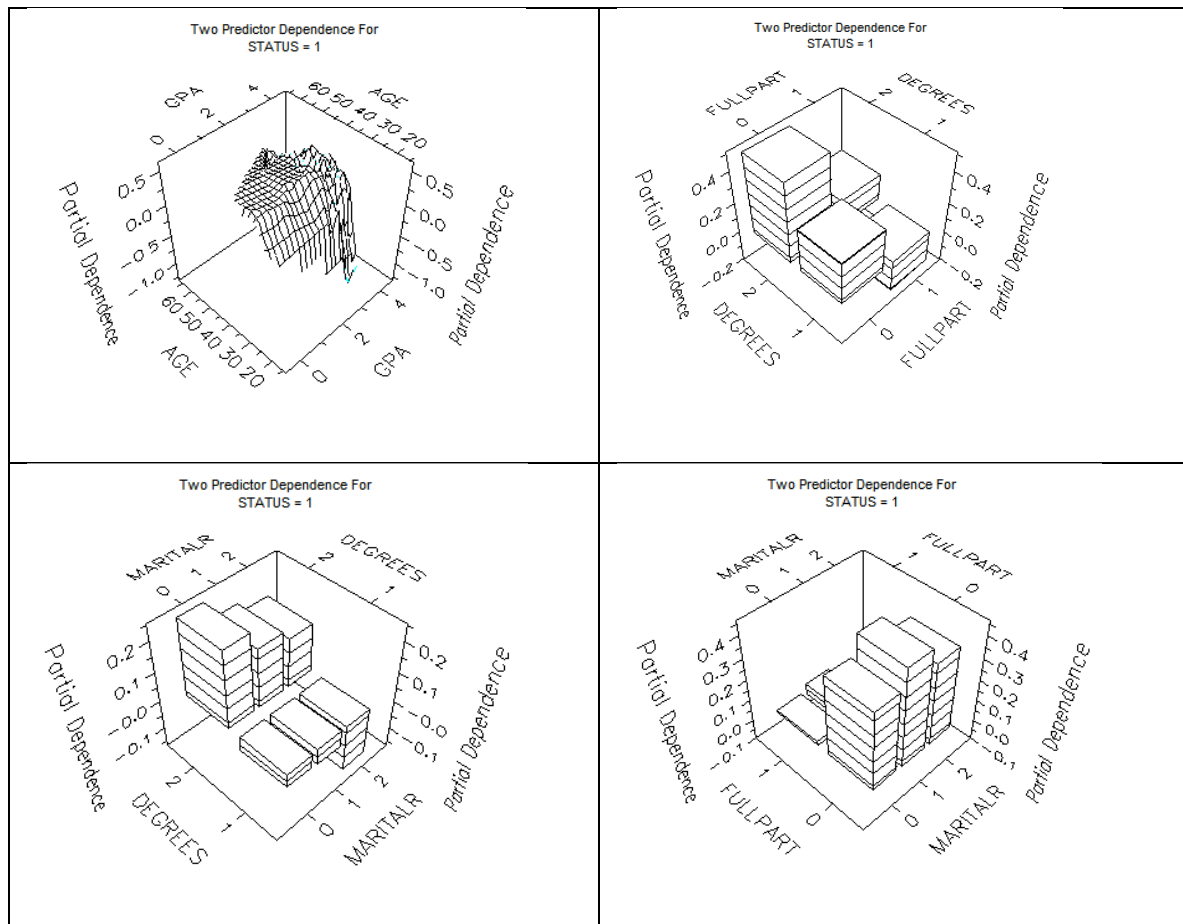


Figure 5. Partial Effect of Degree Program, Status and Marital status on the Dropout Rate (controlling for other predictors)



**Figure 6.** Interaction effects among pairs of variables.

The TreeNet<sup>®</sup> analysis indicates that the risk of attrition increases with age and then declines and levels off for ages over about 40. The risk of attrition is high but relatively constant for entering GPA values less than about 3.0, after which it decreases to a new level for GPAs above about 3.2. The TreeNet<sup>®</sup> analysis confirms the interpretations for the impact of degree program, full-time and part-time status and marital status. In addition, the TreeNet<sup>®</sup> analysis suggests a broader set of interactions between the variables than the other analyses.

### Summary and Conclusions

Our objective in this article was to identify factors that could assist business schools to maximise student retention. The models we utilised in this research provide a means to identify at-risk students, allowing the business schools to develop proper intervention strategies to prevent premature dropout. One of the unique aspects of this problem is the inclusion of part-time students who may stretch their degree program out over a number of years. The inclusion of part-time students, which can make up a significant portion of the business school graduate student enrolment, renders a cohort analysis inappropriate. For practical reasons, we chose to consider only variables that are typically available from either a student's application or attendance records. We employed survival analysis, decision trees and TreeNet<sup>®</sup> in an attempt to accommodate some of the unique aspects of the data. We have not found these methodologies represented in the literature on student attrition. Survival analysis explicitly considers time to estimate the risk of a student dropping out of the program. Decision trees provide a cross-sectional alternative to assess factors important to

predict whether a student drops out. Finally, TreeNet<sup>®</sup> analysis can provide a richer view of the relationships of variables used in the decision trees.

Key findings of our research include the following:

- Full-time students have a higher risk of dropping out.
- The risk of dropping out is higher for married students.
- Students with higher entry GPA have a lower risk of dropping out.
- Students with lower cumulative GPA in school are more likely to drop out.
- Older students are more likely to drop out.
- For MS students the risk of attrition goes up with age.

Taking the results of the different analyses, degree status, marital status, entering GPA and cumulative GPA may be factors that could help identify students at risk of dropping out. Degree program and age may have a more complicated relationship with risk of attrition. This information could provide the foundation to screen for high-risk students. Once identified, outreach programs can be implemented to improve student retention. We note at this point the following caveats to our analysis. Because candidates tend to graduate faster from MS programs than MBA programs, MS candidates constitute a majority in the full dataset but a minority in the analysis dataset. For similar reasons, relative to the full dataset, part-time students are overrepresented in our analysis dataset. It is also important to note that our analysis purposely excludes students who graduate. Thus, it would be important to ensure that any intervention based on these results does not have a negative impact on this desirable outcome.

## References

- Astin, A.W. (1997). How 'good' is your institution's retention rate? *Research in Higher Education*, 38(6), 647–658.
- Booth, A.L., & Satchell, S.E. (1995). The hazards of doing a PhD: An analysis of completion and withdrawal rates of British PhD students in the 1980s. *Journal of the Royal Statistical Society Series A*, 158(2), 297–318.
- Braunstein, A.W., Lesser, M., & Pescatrice, D.R. (2006). The business of freshmen student retention: Financial, institutional, and external factors. *Journal of Business & Economic Studies*, 12(2), 33–53.
- Cleves, M.A., Gould, W.W., & Gutierrez, R.G. (2004). *An introduction to survival analysis using Stata*. College Station, TX: Stata Press.
- DeShields, O.W. Jr., Kara, A., & Kaynak, E. (2005). Determinants of business student satisfaction and retention in higher education: Applying Herzberg's two-factor theory. *International Journal of Educational Management*, 19(2), 128–139.
- Druzdzel, M.J., & Glymour, C. (1994). Application of the TETRAD II program to the study of student retention in U.S. colleges. In *Proceedings of the AAAI-94 Workshop on Knowledge Discovery in Databases (KDD-94)*, 419–430, Seattle, WA.
- Friedman, J. (1999). *Greedy function approximation: A gradient boosting machine*. Retrieved from <http://www.salfordsystems.com/doc/GreedyFuncApproxSS.pdf>

- 
- Friga, P.N., Bettis, R.A. and Sullivan, R.S. (2003). Changes in graduate management education and new business school strategies for the 21<sup>st</sup> century. *Academy of Management Learning and Education*, 2 (3), 233–249.
- Hammond, K., Webster, R.L., & Harmon, H.A. (2006). Market orientation, Top management emphasis, and performance within university school of business: Implications for universities. *Journal of Marketing Theory and Practice*, 14(1), 69–86.
- Johnes, G., & McNabb, R. (2004). Never give up on the good times: Student attrition in the UK. *Oxford Bulletin of Economics and Statistics*, 66(1), 23–47.
- Hosmer, D.W. Jr., & Lemeshow, S. (1999). *Applied survival analysis: Regression modeling of time to event data*. New York: Wiley.
- Lu, J. (2002). *Predicting customer churn in the telecommunications industry*. Paper 114–27, SUGI 27, Retrieved from <http://www2.sas.com/proceedings/sugi27/p114-27.pdf>
- Marcus, R.D. (1989). Freshmen retention rates at U.S. private colleges: Results from aggregate data. *Journal of Economic and Social Measurement*, 15(1), 37–55.
- Ott, M.D., Markewich, T.S., & Ochsner, N.L. (1984). Logit analysis of graduate student retention. *Research in Higher Education*, 21(4), 439–460.
- Rust, R., Zeithaml, V., & Lemone, K. (2000). *driving customer equity: how customer lifetime value is reshaping corporate strategy*. New York: The Free Press.
- Salford Systems. (2010). *TreeNet® overview*. Retrieved from <http://salford-systems.com/products/treenet/overview.html>
- Sanders, L., & Burton, J.D. (1996). From retention to satisfaction: New outcomes for assessing the freshman experience. *Research in Higher Education*, 37(5), 555–567.
- Stock, W.A., Finegan, T.A., & Siegfried, J.J. (2006). Attrition in economics Ph.D. programs. *AEA Papers and Proceedings*, 96(2), 458–466.

---

## Appendix

### *Technical Details*

#### **Survival Analysis**

Following the variable selection process suggested by Hosmer and Lemeshow (1999), we performed a Kaplan–Meier survival analysis for each nominal covariate to test whether each covariate yields significantly different ( $p \leq .01$ ) survival functions for each level of the nominal covariate. Gender does not exhibit a significant difference ( $p > .4$ ) between the two survival functions and is dropped from further analysis. A Cox proportional hazards model was used to show that both age and GPA are significant continuous predictors.

#### **Full Cox Proportional Hazards Model Development**

Subsequent to an individual assessment of each variable, a Cox proportional hazards model was estimated using all the variables under consideration, other than gender. The model was first estimated including only the main effects for the predictors. Several of the variables no longer indicated a significant effect and were deleted from the model one by one. In particular, the coefficients for ethnicity, citizenship and visa type were not significantly different from zero ( $p > .05$ ) and were dropped from the model. At each deletion, the change in the log-likelihood of the fit after deleting the insignificant variable was insignificant ( $p > .05$ ) and the changes in the remaining coefficients were minimal, indicating the stability of the resultant model.

Using the remaining significant variables (degree program, part-time/full-time status, marital status, age and GPA) a Cox proportional hazards model was estimated using all main effects and all first-order interactions. Only the interaction between age and degree program was significant ( $p < .05$ ) and the remaining interaction terms were removed. A final model was fitted yielding a log-likelihood value of 324.6 with 6 degrees of freedom ( $p = .00$ ) indicating that the model as a whole is significant (note that the null hypothesis is that all the  $\beta$ s are zero).

Partial residual plots indicate that the assumption of proportionality of hazards (see for example Hosmer & Lemeshow, 1999) was clearly satisfied for degree program, marital status, age, GPA and the interaction of degree program with age. The partial residuals associated with status show some deviation from linearity with length of time in the program but the deviation was not considered severe enough to limit the interpretation of the results.

#### **Extended Cox Model for Time-Dependent Variables**

In the extended Cox model we considered a dynamic factor: cumulative GPA. We define this time-dependent variable as a student's cumulative GPA computed immediately after the individual's last registered term. If the student was absent for several terms before dropping out, the cumulative GPA for the absent terms is recorded as that of the last registered term for that student. In addition, if the student has dropped out, the cumulative GPA after the last registered term is recorded as 0.

Besides the students' cumulative GPA, we also use the same variables included in the Cox proportional hazards model (degree program, status, marital status, age and GPA). This model is estimated by using all main effects and all first order interactions. Again, only the interaction between age and degree program is significant ( $p < .05$ ), and the remaining interaction terms are removed.

A final model was estimated yielding a chi square value of 1343.097 with 7 degrees of freedom ( $p = .00$ ) indicating that the model as a whole is significant.

### Decision Trees

We used the variable indicating whether a student was active or not as a target variable and the full set of variables initially included in the survival analysis as predictors: degree program, status, marital status, citizenship, visa, ethnicity, gender, GPA and age. Recalling the fact that the percentage of dropouts in the dataset is high, we are interested in finding nodes in the tree representing a higher percentage of dropouts than the overall percentage in the dataset (about 79%).

The first decision tree was generated using the CHAID method and the results are displayed in Figure 1. At each stage of the analysis, the CHAID algorithm considers all possible splits of predictor variables and determines the one that best discriminates the target variable values on the basis of associated chi-squared statistics. The second decision tree was obtained by way of the C&RT algorithm and the results are displayed in Figure 2. The C&RT algorithm can be summarised as follows:

- C&RT divides a dataset into segments with as little variability as possible in the dependent variable; for a categorical target variable, the variability is evaluated by a measure of nonhomogeneity such as for example the Gini coefficient.
- C&RT uses the independent variables, continuous or categorical, to split the sample. It uses binary splits such as  $X \leq C$ , where  $X$  is any continuous independent variable,  $C$  any value taken by  $X$ , or for instance  $X = 2,6$  versus  $X = 1,3,4,5$ , where  $X$  is categorical with values coded 1–6.
- Among all possible candidate splits, C&RT selects that split which minimises the total variability in the two new nodes.
- C&RT creates a list of trees from the smallest tree with only one node to the largest tree with as many nodes as observations in the dataset, and then selects the tree that predicts the dependent variable best on an independent test sample.

### TreeNet<sup>®</sup>

The objective of the TreeNet<sup>®</sup> analysis is to obtain a more precise understanding of any nonlinear relationships between the propensity to drop out of graduate school and the predictors, as well as of any interaction among the predictors. Indeed, TreeNet<sup>®</sup> is recommended as a tool to be used once the main predictors have been identified and data quality issues treated (Salford Systems, 2009).

While directly interpreting a TreeNet model is quite complicated, graphs that arise from the procedure display the partial impact of the each predictor separately, as well as the impact of pairs of predictors on the target variable (Salford Systems, 2009). For a description of the TreeNet methodology, we refer the reader to Salford Systems (2009) and Friedman (1999).

The TreeNet<sup>®</sup> model we built yielded an approximation with a sum of 196 trees, and derived the following relative importance of the predictors; the most important variable is assigned the value of 100 and used as a reference point for the other variables.