Variable Dictionary for Student Data and Measures

Table 1: Student information from university's registrar reporting system.

Variable	Description
TERM	The Semester's Year and Term
TERM_GPA	The student's overall GPA (on a four point scale) for courses taken during the term
TERM_ATTEMPTED_HRS	The number of credit-hours the student registered for at the beginning of the term
TERM_PASSED_HRS	The number of credit-hours passed for the term
TERM_QUALITY_POINTS	A product of the number of course credit-hours and the point-assigned letter grade. For example, a student gets 12 points if they receive an A (four grade-points) in a three credit-hour course
PROG_GPA	Overall Program GPA (on a four point scale)

Table 2: Measures produced for each student that capture changes during an *event*.

Measure	Description
Num_Semem_To	The number of semesters for the event.
AvgGPAChg_To	The average change in the TERM_GPA during the event.
AvgProGpa_To	The average PROG_GPA during the event.
Base_GPA	The student's TERM_GPA in their very first semester
AvgDelta_AttPass_To	The average difference between the TERM_ATTEMPTED_HRS and the TERM_PASSED_HRS during the event.
AvgDiffPts_To	The average difference between the maximum points possible (TERM_PASSED_HRS * 4) and the points earned (TERM_QUALITY_POINTS) during the event.

Mathematical Models and Hypotheses

For this study, Num_Semem_To is our response variable, *y*, representing the number of semesters for the *event*.

For these data, Num_Semem_To equals the number of student records having a TERM less than or equal to 201205. We construct the additional average levels using the following models:

$$\boldsymbol{AvgGPAChg\_To} = \begin{cases} 0 & y \leq 1 \\ \dfrac{\sum_{i=1}^{y-1}(TERM\_GPA_{i+1} - TERM\_GPA_i\ )}{y-\ 1} & y > 1 \end{cases}$$

$$\boldsymbol{AvgProGpa\_To} = \begin{cases} 0 & y = 0 \\ \dfrac{\sum_{i=1}^{y}\ PROG\_GPA_i}{y} & y > 0 \end{cases}$$

$$\boldsymbol{AvgDelta\_AttPass\_To} = \begin{cases} 0 & y = 0 \\ \dfrac{\sum_{i=1}^{y}(TERM\_ATTEMPTED\_HRS_i - TERM\_PASSED\_HRS_i\ )}{y} & y > 0 \end{cases}$$

$$\boldsymbol{AvgDiffPts\_To} = \begin{cases} 0 & y = 0 \\ \frac{\sum_{i=1}^{y} (TERM\_PASSED\_HRS_i * 4) - TERM\_QUALITY\_POINTS_i}{y} & y > 0 \end{cases}$$

Using these average levels, along with the demographics, the linear regression model that we use to examine the statistical significance of the measures is represented by:

$$\boldsymbol{Num\_Semem\_To} = \boldsymbol{f}(\boldsymbol{AvgGPAChg\_To}, \boldsymbol{Base\_GPA}, \boldsymbol{AvgProGpa\_To}, \boldsymbol{AvgDelta\_AttPass\_To}, \boldsymbol{AvgDiffPts\_To}, \boldsymbol{Age}, \boldsymbol{Gender}, \boldsymbol{Race}) \ (1)$$

In this work, we create an alternative measure using Num_Semem_To that indicates whether a student has a relatively low, or relatively high risk associated with being retained by the institution. We assign values to this measure, Retention_Risk, using the following model:

$$\boldsymbol{Retention\_Risk} = \begin{cases} Low & Num\_Semem\_To \leq 3 \\ High & Num\_Semem\_To > 3 \end{cases}$$

This work also examines the significance of the measures listed in Table 2 using a Logistic regression model with Retention_Risk as the nominal response variable. Thus, a second model used in this study to explore the statistical significance of the measures we create is represented by:

$$\boldsymbol{Retention\_Risk} = \boldsymbol{f}(\boldsymbol{AvgGPAChg\_To}, \boldsymbol{Base\_GPA}, \boldsymbol{AvgProGpa\_To}, \boldsymbol{AvgDelta\_AttPass\_To}, \boldsymbol{AvgDiffPts\_To}, \boldsymbol{Age}, \boldsymbol{Gender}, \boldsymbol{Race}) \ (2)$$

Logistic regression uses a linear combination of the inputs to generate a logit score that represents the log of the odds of Retention_Risk occuring. Parameter estimates are obtained using maximum likelihood estimation and prediction estimates are produced using a logistic function, which is the inverse of the logit function.

Our key research objective is to examine whether or not our student measures appearing in Table 2 will appear as statistically significant predictors for the respective response variables in Equations (1) and (2). The p-value is used to determine statistical significance.