

Predicting Student Loan Default for the University of Texas at Austin

By *Elizabeth Herr and Larry Burt*

Elizabeth Herr is senior statistician for Noel-Levitz in Denver, Colorado.

Larry Burt is associate vice president of student affairs and financial aid director for The University of Texas at Austin.

During spring 2001, Noel-Levitz created a student loan default model for the University of Texas at Austin (UT Austin). The goal of this project was to identify students most likely to default, to identify as risk elements those characteristics that contributed to student loan default, and to use these risk elements to plan and implement targeted, pro-active interventions to prevent student loan default. UT Austin supplied academic data for the project, and the student loan guarantor Texas Guaranteed Student Loan Corporation (TG) provided the data about borrowers from UT Austin who entered repayment between 1996 and 1999. Results showed that student program completion, persistence, and success were strong predictors of student loan default, as were race/ethnicity, gender, and the school of enrollment at UT Austin. These results emphasize the role of student success and graduation in eventual loan repayment. Interventions that focus on student persistence and academic success were seen as the primary actions needed to help prevent student loan default.

Over the past decade, total aid to students to finance higher education has increased by 117 % (College Board, 2002). In 2002-2003, more than \$105 billion in total financial aid was provided from all sources (College Board, 2003). During the 1990s, the amount of grant aid doubled, while loan aid tripled. The share of grants decreased from 50% of total aid in 1991-1992 to 40% in 2001-2002, while the proportion of aid from loans increased from 47% to 54%. Graduate students use three times as much loan aid as grant aid (College Board, 2002). In 2002-03, federal loans comprised 45% of total aid, amounting to \$47.7 billion (College Board, 2002 & 2003). Overall, 29% of all undergraduates borrowed from some source to help finance their postsecondary education in 1999-2000 (Clinedinst, Cunningham, & Merisotis, 2003).

Of the borrowers with Stafford Loans and/or Supplemental Loans for Students (SLS), undergraduates at two-year public colleges were the least likely to borrow (6%), followed by student borrowers at public four-year schools (35%), private not-for-profit four-year schools (43%), and private for-profit (proprietary) schools at 50% (Berkner, 2000).

Researchers have carefully examined the increasing loan exposure of students over the past 20 years. Studies range from concerns over the overall debt burden facing students after college to several detailed studies about the causes of student loan default. Indebtedness studies have generally concluded that debt burdens are not too high for graduating students and do not

Program completion, student success, and persistence are among the strongest predictors of loan default in virtually all studies.

postpone major purchases such as houses and cars, or affect life decisions, such as marriage. The students with the most difficulties were those who did not obtain their degree or faced challenges such as unemployment, divorce, additional dependents, or incarceration (Greiner, 1996; Texas Guaranteed, 1998a; Choy, 2000; Choy & Li, 2005).

Student loan default has received much attention, especially since the early 1990s, when default rates reached extremely high levels, particularly at proprietary schools. Since then, the average school default rate has declined from a high of 22.4% in 1990 to its lowest level to date, 5.2% in 2002. Nevertheless, student loan default is a serious issue for borrowers, schools, lenders, and guarantors.

Prior studies on the causes of student loan default have focused on the roles of individual student background characteristics versus the characteristics of the schools in which these students had enrolled. Generally, individual student background characteristics outweighed school characteristics as predictive variables. Particularly, race emerged as a highly predictive variable, with Black students being at higher risk of student loan default than Asian or White non-Hispanic students (Wilms, Moore & Bolus, 1987; Knapp & Seaks, 1992; Dynarski, 1994; Flint, 1994; Volkwein & Szelest, 1995; Flint, 1997; Woo, 2002).

Some cross-sectional studies that have combined data from many different schools and school types have found some connection between attending a proprietary school and an increased risk of loan default (Wilms, Moore & Bolus, 1987; Dynarski, 1994; Texas Guaranteed, 1998b), while in other studies, school type did not emerge as significant (Woo, 2002). Proprietary schools appeared as a significant risk factor, in part due to their own lending practices and their tendency to enroll students from low-income backgrounds. An additional factor may be that many studies examined proprietary schools during the early 1990s, before a number of proprietary schools with extremely high default rates were excluded from the federal student loan program.

Finally, program completion, student success, and persistence are among the strongest predictors of loan default in virtually all studies (Wilms, Moore & Bolus, 1987; Knapp & Seaks, 1992; Flint, 1994 & 1997; Volkwein & Szelest, 1995; Texas Guaranteed, 1998a, 1998b; Woo, 2002; Gladieux & Perna, 2005).

This study examines the risk factors for student loan default for borrowers who had attended the University of Texas at Austin (UT Austin) and entered repayment between 1996 and 1999. In recent years, UT Austin has had relatively low student loan default rates, ranging from 6.9% in 1997 to 3.0% in 2002. The median indebtedness for students for academic year 1996-1997 was \$13,993 (Texas Guaranteed, 1998a) and rose to \$18,856 in 2001-2002. Despite the overall low default rate, stu-

dent loan default prevention continues to be an important goal at UT Austin. The intent of this study is to help prevent future defaults by identifying possible interventions while the students are still enrolled. This emphasis on identifying potential points of intervention sets this study apart from other studies of its kind.

This study resulted in a predictive model that included only those variables that could be used to formulate proactive student interventions. This model was designed to allow the institution to look at the predictors very early in the students' undergraduate careers. When variables signaling a higher propensity for default were present, an appropriate level of intervention could be applied. To that end, UT Austin formulated a response plan to help prevent defaults. School officials hoped that the presence of a statistical analysis would help in developing a response that would cross several departmental lines at UT Austin.

Data

Repayer and Defaulter Data File Creation

The data for this study were derived from a source file generated by Texas Guaranteed Student Loan Corporation (TG), the National Student Loan Database System (NSLDS), and UT Austin. The files provided by TG and NSLDS included information about the students in repayment or default from January 1996 through December 1999, and all loans for these students, except Parent Loans for Undergraduate Students (PLUS) and consolidation loans. This data file contained information on 89,994 loan records for 23,418 students. The loan record data was collapsed to the student level, in each case keeping only the last loan status for each loan. This loan status could then be classified as "defaulted" or "other." The loan status "defaulted" became the dependent variable for the study.

Academic and demographic information from UT Austin was appended to the loan default data. The UT Austin data file contained information on students' demographic characteristics; parents' information; students' income and other economic characteristics; and admissions data such as high school records, degree sought, credit hours taken, grade point average (GPA), and transfer information. The original data file contained more than 200 data fields. The UT Austin file contained 23,407 records, all of which were matched to the loan default data. (Eleven borrowers from the loan record file did not match the UT Austin data and were not included in the study.) Of the 23,407 in the final modeling file, 1,306, or 5.58%, showed a final status of default. This rate is slightly higher than official average loan default rates for UT Austin since 1997, which are shown in Table 1. This reflects in part the difference between the official "cohort default rate" versus the proportion of borrowers that ultimately default but not within the period in which the default cohort is calculated.

Table 1
University of Texas at Austin Loan Default Rates,
1997-2002

Cohort Year	Default Rate	Borrowers	Defaulters
2002	3.0%	6,538	198
2001	4.0%	6,771	277
2000	3.8%	7,057	269
1999	3.5%	7,066	254
1998	4.8%	6,434	314
1997	6.9%	6,322	438
Total/Average	4.7%	26,879	1,275

Source: NSLDS Default Rate Tables, 2001, 2002, and 2003.

Methodology

This project comprised two distinct parts: an investigative research portion and a data mining portion. While based on the same data set, different methodologies were used for each portion. For both parts, logistic regressions were estimated using the likelihood of default as the dependent variable. The differences in the methodologies pertained to variable selection and model testing procedures.

Research Methodology

The pure research portion of the project consisted of systematically testing the various groups of academic and demographic data to see which variables were predictive of eventual loan default. The input data represented different aspects of students' backgrounds. In order to test the relative contribution of each set of variables, the data were divided into thematic groups, each group focusing on one aspect of the students' background and experience. Data was entered into the series of logistic regressions incrementally in six different blocks: demographic and background data, high school information, degree and major data, credit hour information, transfer information, and any available financial data.

The regressions used the full set of data, and the predictive power of the model was ascertained by looking at the regression chi-square, the pseudo *R*-squared, and the statistical significance of individual variables. All variables entered into the regressions were tested for their direct correlation with the dependent variable and their mutual intercorrelation. Variables displaying a high degree of intercorrelation were not entered into the regression together, keeping the variable with the higher correlation to the dependent variable in the research regression.

Data Mining Technology

Data mining is a modeling technology that tries to create the model that best predicts a certain outcome. In this case, the goal was to find the model that best predicted which borrowers

were most likely to default, and that best separated the borrowers into two groups: defaulters and repayers. Again, a logistic regression was used to predict the likelihood of default. In this case, the data set was divided into two halves. The first half of the data was used to build the model, while the other half, or holdout sample, was used to score the data with the new model. Since outcomes are known in the holdout sample, it is then possible to validate how well the model predicted correctly, and how well the model was able to separate defaulters from non-defaulters by the assigned model score. This methodology tests the predictive power of each possible model on an independent data set at each point in the modeling process.

This process does not rely on entering the data into the regression based on theoretical or thematic grounds. The original variable selection depends on the correlation between each variable and the final outcome, taking care that variables that are too intercorrelated are not entered into the regression together. Building a model using this technology is an iterative process in which the final number of variables depends on the mix of variables that best predicts the outcome. Over-fitting the model by including many variables that are statistically significant, but contribute only marginally to the estimated outcome, is prevented by choosing the model with the fewest variables that result in the best outcome when scoring the holdout sample. It is expected that the final model produced by the data mining process is similar in variable content to the final model produced by the more thematic research methodology.

Data Limitations

Much of the sample available had a high percentage of missing data. While it is customary in academic research to eliminate all observations with missing data, this was not done in this project. In keeping with data mining conventions, missing data was imputed wherever possible by substituting the mean response or data value for observations with missing data. Using this approach, all observations were kept in the initial modeling process, allowing for investigation of the maximum amount of available data characteristics. Ultimately, however, variables with more than 90% imputed data were eliminated from the modeling process. This affected data fields such as student honors, joint degrees, major codes 3-7, number of dependents, and surprisingly, high school GPA. The final modeling regressions included only those variables with the lowest percentage of missing values.

General Treatment of Variables

Data used in this project were either numeric or categorical. Numeric variables, whether continuous, ordinal, or binary, were entered into the regression in their original form. In some cases, continuous information was also collected into a binary flag that showed the presence or absence of a certain characteristic. For

example, the variable “Transfer Flag” had a value of “1” for all students who had transfer hours greater than zero, and a value of “0” for students who had no transfer work. Students with no data in that particular field received a missing value. Missing values were substituted with the mean value of that variable, a process which does not bias the estimated coefficients. The danger of imputing data is that the missing values are not random, but show a systematic bias. While it is possible to test for this by creating flags that designate missing data for a particular variable, the authors chose to exclude all variables with a high percentage of missing data. In this data set, missing data was deemed to be more of a symptom of data collection or data translation over a long series of years than attributes of the borrower. The final model used variables with minimum percentages of imputed missing data.

Categorical data, such as race/ethnicity or geographic variables (e.g., state of residence) are most often handled by creating one binary dummy variable, or flag, for each category. In the case of variables with a large number of categories, this can lead to an unmanageable number of dummy variables. To avoid this, an alternative treatment of categorical variables is sometimes used. In this treatment, referred to as “classifying” the variable, the numeric response frequency is substituted for the actual category. The result is a single numeric variable that may have fewer response levels, but that keeps the information for each category within one variable. For example, White, non-Hispanic borrowers had an average default rate of 4.61% and African-American borrowers had an average default rate of 12.26%. The classification process substituted the value 0.0461 for all White borrowers and the value of 0.1226 for all African-American borrowers. Categories with a small number of observations are excluded from this process and are instead assigned a missing value. These missing categories then receive the mean response frequency for the file. This avoids the effects of small numbers and exaggerated response rates in the resulting variable.

If the spread between the default rate of the lowest and highest category is large enough, a classified categorical variable will appear as significant in the regression and have a positive coefficient. In data mining, where the goal is to be able to assign a predictive score to each observation, this process ensures that all categories of a variable are weighted in proportion to the risk arising from that particular characteristic.

If, for example, the race/ethnicity variable appears as significant in the regression, this means that there are strong differences in the average default rates of different ethnic groups. Referring back to a table with average default rates for each ethnic group then shows which groups are at highest risk of default. While a dummy variable for each ethnic group would most likely also identify the group with the highest risk as a

significant variable, the differential information on other ethnic groups would be lost.

The classification process is most useful for variables with many response levels, such as state of residence. While using dummy variables for each state would identify one or more states as having students most at risk for loan default, using the variable in its classified version would indicate that the differential average loan default rates between states is significant. Again, referring to a table showing the average loan default rates for each state would identify those states that have above-average loan default rates. In the scoring process, the average default rates for all states would be included and add a differential weight to each individual score.

Borrower Profile

Of the 23,407 borrowers in the sample, approximately half (50.2%) were male, and the average current age was 30. The majority of borrowers were White, non-Hispanics (66%), followed

Table 2
Means of Numeric Variables

Variable	Mean	Standard Deviation	Minimum	Maximum	Percent Missing
Age	30.113	5.759	20.000	66.000	0.00
Disability	0.018	0.088	0.000	1.000	60.67
Armed forces	0.046	0.138	0.000	1.000	56.40
Sex (male=1, female=0)	0.502	0.500	0.000	1.000	0.00
Parents' aggregated income	\$22,154.66	\$32,992.96	0.000	\$99,999.00	49.60
High school class rank-categorized	0.056	0.018	0.039	0.120	0.00
ACT Composite Score	24.260	1.587	11.000	35.000	82.04
SAT Quantitative Score	584.101	56.438	300.000	800.000	51.73
SAT Verbal Score	579.135	62.454	230.000	800.000	51.73
Current GPA	2.927	0.734	0.040	4.000	7.82
Credit hours failed > 0	0.349	0.477	0.000	1.000	0.00
Academic probation flag	0.276	0.447	0.000	1.000	73.90
Credit hours incomplete > 0	0.036	0.186	0.000	1.000	0.00
Credit hours passed	75.742	47.752	0.000	277.000	0.00
Transfer flag	0.612	0.487	0.000	1.000	38.79
Transfer GPA	1.009	1.516	0.000	4.800	69.18
Graduate studies flag	0.284	0.451	0.000	1.000	71.62
Adjusted gross income	\$7,335.79	\$14,150.59	0.000	\$99,999.00	28.56
Taxes paid	\$632.15	\$1,718.75	0.000	\$32,000.00	54.01
Last amount collected	\$3,503.29	\$3,184.01	0.000	\$12,964.00	40.30
Net guarantee	\$4,018.17	\$3,378.86	0.000	\$93,221.00	0.000

Note: Dollar amounts are rounded to the next cent.

by Hispanics (19%), Asian-Americans (8.5%), and African-Americans (5.9%). Almost 80% of the borrowers were Texas residents. Approximately 40% of borrowers had a high school rank at or above the 80th percentile. Instead of the total loan amount, the net guarantee amount was included in the data set. The net guarantee amount is the loan amount minus any lender or guarantor fees, making it slightly lower than the actual loan amount. The average net guarantee was \$4,018.17; the average net guarantee for repayers was \$4,034.57 while the average net guarantee for defaulters was \$3,740.66. Other studies have shown that borrowers with lower loan amounts tend to have higher default rates, reflecting early departure and non-completion of degree (Woo, 2002).

Table 2 shows the mean values of all numeric variables submitted to regressions and the percentage of missing values. Table 3 shows loan default frequencies and rates for selected variables.

Research Models and Results

To assess the importance of various groups of variables to the risk of student loan default, we investigated four basic groups of variables: student demographics and parent background; high school academic performance; college degree sought and GPA; and college credit hour information. We also examined transfer hours, graduate studies information, and financial data.

The focus of this study was to identify the stage of a student's educational experience where the school could best intervene to help avoid potential future loan defaults. For example, strong predictors of default coming from the student's background might suggest a need for increased attention to first-generation students. Predictors among high school performance variables might suggest a need for remedial courses, while college GPA and degree predictors might suggest a need to direct the institution's efforts toward student success and degree completion. Although all of these points of student contact with the institution are important, we designed our research model to indicate the most appropriate type and timing of interventions for students at UT Austin.

After the initial regression including student background information, each subsequent regression retains the previous set of variables and adds the new group of variables. As a result, variables that were predictive in the earlier regressions shifted in predictive power and significance as new information was included. The results of the series of regressions, including the data mining regression, appear on a table in the Appendix. The table shows the raw regression coefficient and the *p*-value of those variables with a significance level of 0.05 or lower.

Demographic Data

The demographic variables entered into the first regression included age, race/ethnicity, gender, disability, service in the

Table 3
Frequencies of Selected Variables

Value	Total Number	Percent (%)	Number Defaulted	Default Rate (%)
All	23,407	100.0	1,306	5.58
Gender				
Male	11,749	50.2	810	6.89
Female	11,657	49.8	496	4.25
Race/Ethnicity				
African American	1,378	5.9	169	12.26
Hispanic	4,383	18.7	319	7.28
Native American	118	0.5	8	6.78
Asian American	1,981	8.5	94	4.75
White/non-Hispanic	15,536	66.4	716	4.61
Missing values	9	0.0	0	0.00
Other	2	0.0	0	0.00
Current Age				
20-24	2694	11.5	270	10.02
40+	1765	7.5	125	7.08
25-29	9555	40.8	499	5.22
30-39	9393	40.1	412	4.39
Texas Residency Status				
Texas resident	18,388	78.6	1,155	6.28
Non-Texas resident	3,155	13.5	87	2.76
Foreign resident	5	0.0	0	0.00
Not provided/missing	1,859	7.9	64	3.44
Highest Degree: Father				
High school diploma	281	1.2	23	8.19
Baccalaureate	215	0.9	15	6.98
Associate degree	1,942	8.3	118	6.08
Certification of completion	4,682	20.0	229	4.89
Missing values	16,287	69.6	921	5.65
Highest Degree: Mother				
High school diploma	305	1.3	24	7.87
Baccalaureate	119	0.5	8	6.72
Associate degree	2,744	11.7	156	5.69
Certification of completion	4,003	17.1	209	5.22
Missing values	16,236	69.4	909	5.60
High School Class Rank-Categorized				
25.01 - 50.00 Percent	1,092	4.7	114	10.44
Missing Values	30	0.1	3	10.00
0.01 - 25.00 percent	336	1.4	32	9.52
50.01 - 60.00 percent	863	3.7	78	9.04
60.01 - 70.00 percent	1,331	5.7	98	7.36
70.01 - 80.00 percent	2,151	9.2	149	6.93
80.01 - 90.00 percent	3,542	15.1	216	6.10
90.01 - 100.00 percent	5,778	24.7	256	4.43
Unknown	8,284	35.4	360	4.35
Highest Degree:borrower				
High school diploma	5,058	21.6	801	15.84
Special professional	16	0.1	1	6.25
Baccalaureate	11,592	49.5	397	3.42
Masters degree	4,392	18.8	70	1.59
Doctoral degree	2,349	10.0	37	1.58

Note: Unless otherwise indicated, the categories are sorted from highest to lowest loan default rate.

**Table 3 (cont'd.)
Frequencies of Selected Variables**

Value	Total Number	Percent (%)	Number Defaulted	Default Rate (%)
Highest Class Level				
Freshman	855	3.7	186	21.75
Sophomore	988	4.2	154	15.59
Junior	1,165	5.0	154	13.22
Senior	12,916	55.2	647	5.01
Doctoral	1,673	7.1	55	3.29
Masters	4,274	18.3	91	2.13
Law School	1,527	6.5	19	1.24
Professional	8	0.0	0	0.00
Missing Values	1	0.0	0	0.00
School of Degree #1				
No Degree Attained	6067	25.9	857	14.13
Social Work	179	0.8	10	5.59
Fine Arts	565	2.4	23	4.07
Liberal Arts	4385	18.7	171	3.90
Education	840	3.6	29	3.45
Communication	1456	6.2	40	2.75
Business Administration	1372	5.9	33	2.41
Natural Sciences	1656	7.1	39	2.36
Not provided	393	1.7	8	2.04
Graduate School	2910	12.4	59	2.03
Engineering	1175	5.0	22	1.87
Law School	959	4.1	7	0.73
Graduate Business	1219	5.2	8	0.66
Nursing	231	1.0	0	0.00
Cumulative College GPA				
0.00 - 0.99	372	1.6	86	23.12
1.00 - 1.99	2,085	8.9	391	18.75
2.00 - 2.49	3,374	14.4	320	9.48
2.50 - 2.99	4,515	19.3	213	4.72
Unknown	1,830	7.8	60	3.28
3.00 - 3.49	5,150	22.0	120	2.33
3.50 - 4.00	6,081	26.0	116	1.91
Credit Hours Failed Flag				
Yes: Credit hours failed > 0	8,170	34.9	944	11.55
No: Credit hours failed = 0	15,237	65.1	362	2.38
Financial Need Level				
Independent-single	2,909	12.4	252	8.66
Zero parental contribution	2,620	11.2	217	8.28
Parental contribution: \$1-\$3000	1,810	7.7	128	7.07
Independent-married	1,327	5.7	82	6.18
Parental contribution: > \$3000	7,447	31.8	428	5.75
Z-Missing values	1,890	8.1	98	5.19
Graduate	4,566	19.5	93	2.04
Graduate-married	838	3.6	8	0.95
Dependent/Independent Status				
Dependent	12,406	53.0	798	6.43
Independent	9,805	41.9	449	4.58
Missing values	1,196	5.1	59	4.93

Note: Unless otherwise indicated, the categories are sorted from highest to lowest loan default rate.

**Table 3 (cont'd.)
Frequencies of Selected Variables**

Value	Total Number	Percent (%)	Number Defaulted	Default Rate (%)
Net Guarantee Amount (in order of increasing net guarantee amount)				
\$1-1,500	3,316	14.2	221	6.66
\$1,501-3,000	7,737	33.1	523	6.76
\$3,001-4,500	4,045	17.3	233	5.76
\$4,501-6,000	5,325	22.8	228	4.28
\$6,001-7,500	996	4.3	28	2.81
\$7,501-9,000	1,134	4.8	22	1.94
\$9,001-10,500	440	1.9	17	3.86
\$10,501-12,000	90	0.4	0	0.00
\$12,001-15,000	134	0.6	9	6.72
\$15,001-18,000	50	0.2	2	4.00
\$18,001-21,000	34	0.1	4	11.76
\$21,001-24,000	12	0.1	5	41.67
> \$24,000	93	0.4	14	15.05

armed forces, citizenship, Texas residency status, the highest degree attained by the father and mother, and parents' aggregated income. The initial regression showed that three variables were significant at the $p = 0.001$ level: race/ethnicity, gender, and Texas residency status. Of the different racial/ethnic categories, Blacks and Hispanics were more likely to default than Whites and Asians. This finding is supported by several other studies (Wilms, Moore & Bolus, 1987; Knapp & Seaks, 1992; Dynarksy, 1994; Flint, 1994, 1997; Volkwein & Szelest; 1995; Woo, 2002). In this study, men were more likely to default than women. This result is also upheld in some prior studies (Flint, 1994, 1997; Woo, 2002). Texas residents were more likely to default than non-Texas residents.

Of other student characteristics, the disabilities flag was significant at the $p = 0.05$ level, but this variable had 60% missing data and a low number of students with disabilities. The significance of the parents' aggregated income variable indicated that students whose parents have higher incomes are less likely to default. This result has been found in previous default studies (Wilms, Moore & Bolus, 1987; Knapp & Seaks, 1992; Dynarksy, 1994; Woo, 2002). Of the background variables, only race/ethnicity, gender, and parents' income remained statistically significant as other groups of variables were added to the regression.

The general result of this regression implies that minority students, particularly Blacks and Hispanics, are at a higher risk of default. In addition, students coming from families with lower incomes are also at higher risk. These students might benefit from increased attention from UT Austin in the form of interventions that help students integrate into the campus community and meet the cost of college education.

Interestingly, a higher SAT verbal score was weakly linked to loan default, a result that remained constant across all regressions.

High School Data

The second grouping of data included variables capturing students' high school performance. Unfortunately, high school GPA was not included in the data set for unknown reasons, but high school rank, advanced placement hours, and ACT and SAT test scores were in the data set. Of the high school variables, high school rank, high school College Board code, and the SAT verbal score emerged as statistically significant. Students with lower high school rank were more likely to default. Interestingly, a higher SAT verbal score was weakly linked to loan default, a result that remained constant across all regressions. The counterintuitive results of the SAT verbal score are not easily explained. In this author's experience of retention modeling, the SAT verbal score is often more strongly correlated to student persistence than either the SAT combined score or SAT math score. While the dependent variable of this model is loan default, the result remains puzzling. Neither the SAT math nor SAT combined score entered as significant explanatory variables.

High school College Board code was a categorical variable that was classified. This means that the single variable contained the average loan default rates of all high schools that had more than 12 students attending UT Austin in the modeling file. Generally, high school code can be interpreted as a geographic and academic variable, identifying high schools across Texas and the rest of the country with students who were more likely than average to default.

High school performance and completion have emerged as significant in several cross-institutional studies (Wilms, Moore & Bolus, 1987; Dynarski, 1994; Flint, 1994; Woo, 2002). All studies imply that high school completion and a better high school performance are linked to lower loan default rates. This regression reaffirmed these results, although the particular mix of predictive variables appeared rather unintuitive. For example, it is possible that certain high schools may tend toward strong grade inflation or other characteristics that place their students at increased risk. In the absence of additional information, UT Austin could focus on high school rank as an indicator of eventual loan default.

Degree Completion and GPA Data

Degree completion data emerge as the strongest predictors of loan default status. The most important variables are the highest degree attained, the highest class level reached before leaving UT Austin, and the school at UT Austin from which the student earned the degree. These variables overlap and have some degree of intercorrelation, but were still independent enough to be entered into the regressions together as a group. The data demonstrate that students who earned graduate degrees were the least likely to default. The average default rate of students who received a high school diploma (as opposed to a college

degree) was 15.8%. Borrowers who attained a bachelor's degree had an average loan default rate of 3.4%, and master's and doctoral degree recipients had an average default rate of 1.6% when rounded to the nearest tenth.

Of the students who did not receive a degree, those who left as freshmen were most likely to default (average default rate of 21.75%), followed by sophomores (15.59%) and juniors (13.22%). Students who left as seniors had an average default rate of 5.01%, close to the sample average of 5.58%, while students with graduate or professional degrees had below average default rates. Students who did not receive a degree were more likely to default than any other group of students. These results are echoed by previous studies that find degree completion one of the strongest predictors of loan default (Wilms, Moore & Bolus, 1987; Knapp & Seaks, 1992; Dynarksy, 1994; Flint, 1997; Volkwein & Szelest, 1995; Texas Guaranteed, 1998b; Woo, 2002, Gladieux & Perna, 2005).

Once degree information is added to the model, several variables either gain or lose statistical significance. This happens as the new variables in the model either substitute for, or amplify the effects captured by the other variables. Age was not a statistically significant variable in the first two regressions, but enters the model once degree information is added to the model. The coefficient implies that students who are older are more likely to default, which contradicts findings that students who drop out early, as freshmen, are most likely to default.

One possible interpretation of this result is that the coefficients for the degree variables give too much weight to younger students and that this is compensated for by adding to the default risk of older students through the age variable. Also, older students tend to have other obligations besides paying for college, and these other expenses may account for their higher default tendencies. Table 3 shows that the relationship of age and loan default is not linear, but that students between the ages of 20-24 and over 40 have higher loan default rates than borrowers in their late twenties and thirties. Similarly, high school rank, high school code, and the Texas residency variable lose statistical significance when degree information is included in the regression, and remain insignificant in subsequent analyses.

This regression offered important information for UT Austin in terms of potential student interventions. Student persistence and degree completion emerged as the main variables in this regression. Freshmen persistence, in particular, was important in predicting eventual loan repayment. Enhancing the first-year experience and targeting first-year retention rates appears a worthwhile effort for UT Austin. Based on this data, it would seem that any intervention that helps students persist and succeed in college would substantially lower their risk of loan default.

It would seem that any intervention that helps students persist and succeed in college would substantially lower their risk of loan default.

The inclusion of this level of detail about the student's academic performance is unique to this data set and underscores the effects of student persistence and academic success on future student loan defaults.

College GPA, Hours Failed, Hours Incomplete, Transfer Hours, and Graduate Studies Flag

The data set also contained the students' final cumulative college GPA, the number of hours a student had failed, the number of hours for which the students had received an incomplete grade, and number of hours the student had passed. An additional flag indicated that the student had been placed on academic probation. Of these variables, all but two emerged as highly significant. College GPA was one of the strongest predictors. Students leaving UT Austin with a higher college GPA were less likely to default. Students who had failed any credit hours in college were more likely to default, as were students who had incomplete grades on their academic record. Neither the academic probation flag nor the number of credit hours passed was significant.

The inclusion of this level of detail about the student's academic performance is unique to this data set and underscores the effects of student persistence and academic success on future student loan defaults. In this study, students who had failed any credit hours had an average loan default rate of 11.6% compared with an average default rate of 2.38% for students who had no failed credit hours on their record. This information gives UT Austin another point of early intervention by focusing on students who had any failed credit hours on their record, especially early in their enrollment.

Transfer Hours and Graduate Studies Flag

The presence of transfer credit hours was negatively related to loan default, but a higher transfer GPA had a positive effect on loan default. This result may be due to interactions between variables. Single variable analyses show that students with a higher transfer GPA are less likely to default. While variables that were too highly correlated were omitted from the analysis, this threshold was set rather high (at a Pearson's correlation coefficient of 0.80) and did not preclude some unexpected variable interaction. Completing graduate credit hours was not significant in this regression but gained a low level of significance when student income variables were added.

Adding transfer hours and a graduate studies flag allowed UT Austin to assess the risk level for transfer students and graduate students. The general results upheld that academically strong students and students who complete their undergraduate degree by enrolling in graduate hours are at a lower risk for loan default.

Income and Financial Aid Variables

Of the available income and financial aid data, the amount of taxes paid was highly significant when submitted in combination with the aggregated income variable. When we eliminated the amount of taxes paid from the model, aggregated income

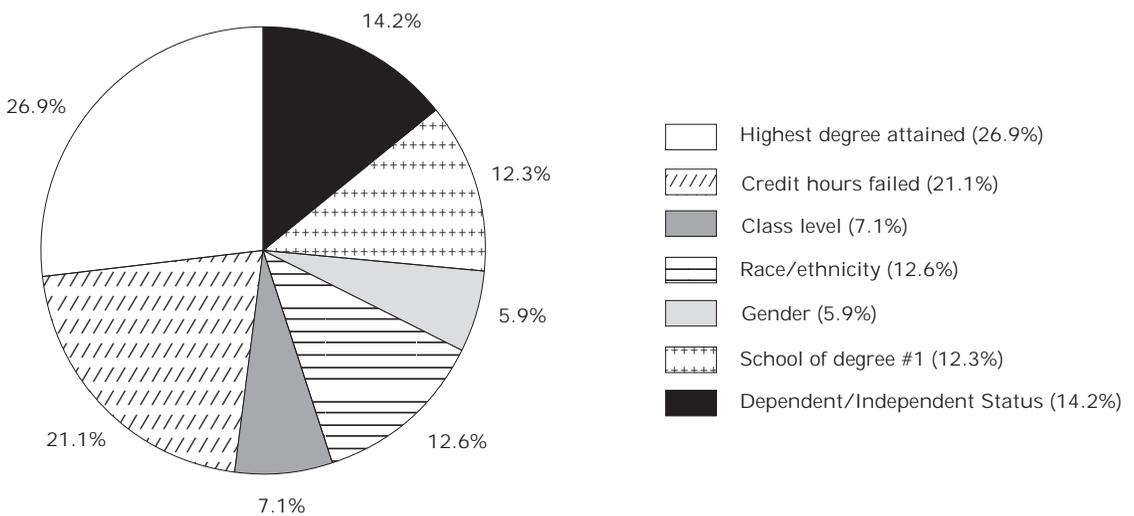
became highly significant. This indicates that borrowers paying taxes—who are also the borrowers with higher incomes—are less likely to default. Students who are employed and have higher incomes have been shown to be at a lesser risk of loan default in other studies (Choy, 2000; Woo, 2002; Choy & Li, 2005). Other income variables tested in the model included financial need, status as financially dependent student, adjusted gross income, and last loan payment amount collected. None of these variables were statistically significant in this regression.

Data Mining Model

The group of variables most highly correlated to student loan default were submitted to the data mining modeling process. This group included 38 variables with correlation coefficients ranging from 0.23 to 0.03. Only variables with a minimum percentage of missing values were considered for this model.

The final model combined the demographic, degree completion, credit hour, and financial variables. Race/ethnicity and gender remained highly significant and accounted for approximately 20% of the variation in default behavior explained by the model. The highest educational degree attained, academic grade level, and school of enrollment variables provided a detailed degree-completion and persistence profile. Taken together, the degree completion variables accounted for more than 50% of the variation in default behavior explained by the model (see Figure). The number of credit hours failed underlined the importance of academic success and explained another 20% of

**Figure
Relative Strength of Model Variables**



The percentages are calculated as the proportion of total variance of the model explained by a particular variable, as measured by the absolute value of the t-statistic for that variable.

borrower behavior, while the financial dependency status variable added a financial aid component to the model. This model was able to predict correctly 76% of the students as defaulters or repayers by the assigned model score. The results of this model echo those of the previous regressions, though in a more efficient model including only seven highly significant variables.

This model provided UT Austin with a succinct profile of potential defaulters that suggested many possible points of intervention spanning a student's educational experience. Student socioeconomic background and possible first-generation student status might be proxied by the race/ethnicity variable. Academic grade level and the credit hours failed emphasized the importance of first-year retention, and the highest degree attained demonstrated the importance of continued student success at all grade levels.

Profile of Student Loan Default

Student loan default can be predicted with limited success from student background variables alone. Both gender and race/ethnicity remain strong predictors throughout all regressions. Based on parents' income variables, students from a higher socioeconomic background are less likely to default. High school performance is important, but only in the absence of college and degree information.

Degree completion and academic success are the strongest predictors of future loan default. Students who completed their degree and have a high college GPA were least likely to default. The earlier a student withdrew from UT Austin, the stronger the likelihood of default. Academic failure—often a precursor to academic withdrawal—also had a strong effect on future default. Failing any credit hours at all increased the possibility of default from 2.38% to 11.55%. These results point to the opportunity of influencing the loan default rate by focusing on student persistence and success at the time a student enrolls at UT Austin.

Of the financial variables, only the amount of taxes paid had any statistically significant influence on default behavior, which suggests that borrowers with higher incomes after leaving school were less likely to default. Other studies with more complete financial data have shown that post-enrollment employment status and higher levels of income lower the likelihood of default and keep the borrower's debt burden at acceptable levels of default (Hansen & Rhodes, 1988; Dynarksy, 1994; Flint, 1997; Volkwein & Szelest; 1995; Choy, 2000; Woo, 2002; Choy & Li, 2005). One way UT Austin could influence student employment is through its alumni network and career counseling.

The data mining model summarized the most salient characteristics that affected student loan default. The goal of the data mining model was to predict future loan defaulters and assign a risk score to each borrower indicating his or her likeli-

hood of default. An additional goal was to find variables that would allow either the loan guarantor or the institution to identify at-risk borrowers as early as possible and take intervention measures to help prevent student loan default. The profile resulting from this model emphasized student background characteristics, degree completion, and the importance of academic success. Because of its comprehensive nature, this was the model best suited for investigating possible student interventions.

Implications of the Models

This study is unusual in that it originated with a student loan guarantor and an institution. The base for this model was cohorts of borrowers who entered loan repayment from 1996 to 1999, and included students from all academic levels and disciplines. While this group of borrowers reflected the loan default issue from the point of view of the loan guarantor, it provided an incomplete picture to the academic institution. The focus on loan default cohorts limited the ability to append complete academic data to all student records and resulted in a data structure that contained many missing values, precluding a truly comprehensive analysis. Nevertheless, the models were able to predict correctly 70% - 79% of the students as defaulters or repayers based on the risk scores derived from the models.

Despite the data limitations, the data show two factors as strongly influencing student loan defaults: student persistence and degree completion. This result provides UT Austin with powerful information about the possibilities of lowering their overall loan default rate and preventing individual loan defaults. Goals for increasing student retention and program completion are well within the scope of UT Austin and can be affected with targeted interventions at the student level. While these interventions will never eliminate default entirely, helping students to succeed will reduce the greatest risk of loan default.

It is possible to take these results one step further and use them to enhance the institution's default reduction efforts. Overall, the estimated models reflect broad trends that emphasize student success as a key factor in reducing defaults. Because the data included students from all academic levels and programs, the model was able to identify the effects of additional years of schooling on loan default rates. Based on the results, it appears that a more direct focus by UT Austin on student retention from freshmen to sophomore year might help the institution to further refine its default prevention efforts.

To achieve this, UT Austin could use the same data mining approach to estimate a freshmen-to-sophomore retention model using all available data for first-year entering students. This model would have the advantage of focusing on an academic cohort rather than a loan default cohort that combines academic years and degrees. The data would be more immediate and the time needed to implement effective policies would be shortened by years. Furthermore, because the model would

be based on more complete and timely data, the predictive factors of this model would signal possible academic interventions tailored to freshmen—the most at-risk group.

Continued Efforts

In the aftermath of the predictive data mining model, UT Austin has both investigated aspects of student retention and sought ways to use the model to plan and implement student interventions, particularly those aimed at students who fail at least one academic class. Several university offices were involved in these efforts. Follow-up information obtained from UT Austin's academic enrichment services (AES) showed that students are most likely to drop out of college during their junior year. In most cases, juniors with low GPAs typically received their first failing grade as early as their first semester. In an effort to boost retention and decrease student loan default, the office of student financial services (OSFS) recently initiated the "Pathway to Progress" (PTP) program. The PTP program combines the efforts of the OSFS, AES, and academic advisors to provide immediate and comprehensive support to freshmen who received at least one failing grade during their first semester. This three-point approach is intended to help reduce financial or academic barriers that may have contributed to the student failing one or more courses.

The PTP program identified approximately 300 aid recipients and divided them into three groups. The first group consisted of students who failed more than one course. These students were required to meet with a representative from OSFS, AES, and an academic advisor. The second group contained Federal Pell Grant recipients with one failing grade. These students met only with a financial aid counselor and an academic advisor. The final group contained non-Pell-eligible students with one failing grade. They were only required to meet with a financial aid counselor. In all cases, the student completed a PTP form where they reported what factors contributed to their failing grade and what they intended to do to improve their academic performance. The students were counseled on using the full extent of services provided by the university.

UT Austin initiated this program late in the spring semester of 2004. Because PTP is designed to be most effective when students are contacted early in spring, the effects are expected to be minimal for fall 2004 freshmen. However, a structure is now in place for productive fall and spring programs. We expect that PTP will expand beyond first-time freshmen to include all grade levels, and anticipate that this program will greatly assist students in obtaining their degrees, which may significantly decrease the likelihood of defaults.

References

- Berkner, L. (2000). *Trends in undergraduate borrowing: Federal student loans in 1989-90, 1992-93 and 1995-96*. (NCES 2000-151). Washington, DC: National Center for Education Statistics, U.S. Department of Education.
- Choy, S. (2000). *Debt burden four years after college*. (NCES 2000-188). Washington, DC: National Center for Education Statistics, U.S. Department of Education.
- Choy, S. & Li, X. (2005). *Debt burden: A comparison of 1992-93 and 1999-2000 bachelor's degree recipients a year after graduating*. (NCES 2005-170). Washington, DC: National Center for Education Statistics, U.S. Department of Education.
- Clinedinst, M. E.; Cunningham, A. F & Merisotis, J. P. (2003). *Characteristics of undergraduate borrowers: 1999-2000*. (NCES 2003-155). Washington, DC: National Center for Education Statistics, U.S. Department of Education.
- College Board, (2002). Trends in student aid 2002. Washington, DC. Author.
- College Board (2003). Trends in student aid 2003. Washington, DC. Author.
- Dynarski, M. (1994). Who defaults on student loans? Findings from the National Postsecondary Student Aid Study. *Economics of Education Review*, 13(1), 55-68.
- Flint, T. A. (1994). The federal student loan default cohort: A case study. *Journal of Student Financial Aid*, 24(1), 13-30.
- Flint, T. A. (1997). Predicting student loan defaults. *Journal of Higher Education*, 68(3), 322-354.
- Gladioux, L. & Perna, L. (2005). *Borrowers who drop out: A neglected aspect of the college student loan trend*. (National Center Report #05-2) The National Center for Public Policy and Higher Education. Online. Available: [<http://www.highereducation.org/reports/borrowing/index.shtml>]
- Greiner, K. (1996). How much student loan debt is too much? *The Journal of Student Financial Aid*, 26(1), 7-16.
- Hansen, W. L. & Rhodes, M. S. (1988). Student debt crisis: Are students incurring excessive debt? *Economics in Education Review*, 7(1), 101-112.
- Hosmer, D. & Lemeshow, S. (2000). Applied Logistic Regression. John Wiley & Sons, Inc.: New York.
- Knapp, L. G. & Seaks, T. G. (1992) An analysis of the probability of loan default on federally guaranteed student loans. *The Review of Economics and Statistics*, 74(3), 404-411.
- Lein, L., Rickards, R., & Webster, J. (1993). Student loan defaulters compared with repayers: A Texas case study. *Journal of Student Financial Aid*, 23(1), 29-39.
- Steiner, M. & Teszler, N. (2003). The characteristics associated with student loan default at Texas A&M University. Austin, TX. Texas Guaranteed Student Loan Corporation.
- Texas Guaranteed (1998a). Education on the installment plan: The rise of student loan indebtedness in Texas. Online. Available: [<http://www.tgslc.org/publications/reports/indebtedness/>]
- Texas Guaranteed (1998b). Student loan defaults in Texas: Yesterday, today and tomorrow. Online. Available: [http://www.tgslc.org/publications/reports/defaults_texas/]
- U.S. Department of Education (16 September 2003). Student loan default rates lowest ever. Online. Available: [<http://www.ed.gov/news/pressreleases/2003/09/09162003.html>]
- Volkwein, J. F. & Cabrera, A. F. (1998). Who defaults on student loans?: The effects of race, class, and gender on borrower behavior in *Condemning Students to Debt: College Loans and Public Policy*. Fossey, R. & Bateman, M. Eds. New York: Teachers College Press.
- Volkwein, J. F. & Szelest, B. P. (1995). Individual and campus characteristics associated with student loan default. *Research in Higher Education*, 36(1), 41-72.
- Wilms, W. W., Moore, R. W. & Bolus, R. E. (1987). Whose fault is default? A study of the impact of student characteristics and institutional practices on Guaranteed Student Loan default rates in California. *Educational Evaluation and Policy Analysis*, 9(1), 41-54.
- Woo, J. (2002). Factors affecting the probability of default: Student loans in California. *Journal of Student Financial Aid*, 32(2), 5-23.

Appendix

Regression Results	Variable Type	Regression 1 Background	Regression 2 High School
Background Variables			
Age	Continuous	-0.00626	-0.00258
Citizenship	Categorical	-1.1544	1.8174
Disability	Dummy	0.5726*	0.5309*
Armed Forces	Dummy	-0.1023	-0.1195
Texas residency status	Categorical	25.5869***	22.7796***
Race/Ethnicity	Categorical	14.0626***	14.1314***
Sex	Dummy	0.5658***	0.5427***
Highest degree: Father	Categorical	6.89	6.0761
Highest degree: Mother	Categorical	-7.9922	-7.2659
Parents' aggregated income	Continuous	-2.98E-06**	-3.32E-06**
High School Variables			
High school class rank (categorized ^a)	Categorical		10.5905***
ACT composite score	Continuous		-0.00122
High school code	Categorical		1.5462*
Advanced placement hours	Categorical		-1.7497
SAT quantitative score	Continuous		-0.00046
SAT verbal score	Continuous		0.00111*
Degree and Enrollment Variables			
Department or school 1	Categorical		
Class	Categorical		
Highest degree attained	Categorical		
Degree #1	Categorical		
Degree major #1	Categorical		
School of degree #1	Categorical		
GPA and Credit Hour Data			
Current GPA	Continuous		
Credit hours failed >0	Dummy		
Credit hours failed	Continuous		
Academic probation flag	Dummy		
Credit hours incomplete >0	Dummy		
Credit hours passed	Continuous		

Regression 3 Degree Info	Regression 4 GPA/Hours	Regression 5 Transfer/Grad	Regression 6 Financial	Regression 7 Data Mining
Background Variables				
0.0488***	0.0526***	0.0503***	0.0554***	
0.9951	2.4216	3.663	3.23	
0.1141	0.1592	0.1747	0.1553	
-0.2093	-0.2156	-0.224	-0.2301	
5.5641	2.4994	6.0384	5.9163	
12.0747***	9.3502***	9.1119***	9.1493***	10.15089***
0.4483***	0.3262***	0.3066***	0.2971***	0.2293**
-3.8356	-5.1905	-5.3847	-4.6895	
-5.8318	-8.2483	-7.9245	-8.1708	
-6.22E-06***	-6.7E-06***	-6.42E-06***	-5.3E-06**	
High School Variables				
0.6001	-0.3213	-0.1855	-0.0384	
0.00181	0.00265	0.00265	0.00307	
0.7976	0.7242	0.733	0.6582	
-4.256	-8.4354	-7.8677	-8.1886	
0.0006	0.0007	0.000807	0.000795	
0.00114*	0.00142**	0.00112*	0.00111*	
Degree and Enrollment Variables				
7.4317***	4.4845*	6.107**	5.6152**	
3.4626***	2.8931**	1.8685*	1.6348	2.5715**
13.02***	9.8868***	10.4094***	10.1733***	11.0232***
5.7079	5.2174	6.1027	5.8365	
0.3798	0.3335	0.3159	0.2347	
13.8177**	11.7674**	13.2788**	12.1557**	23.7350***
GPA and Credit Hour Data				
	-0.3523***	-0.391***	-0.3899***	
	0.5653***	0.5857***	0.5754***	
				0.0369***
	0.0831	0.0674	0.0652	
	1.0785***	0.9591***	0.9307***	
	0.00164	0.00239*	0.00187	

Appendix (cont'd.)

Regression Results	Variable Type	Regression 1 Background	Regression 2 High School
Transfer and Graduate Studies Data			
E101–Transfer flag	Dummy		
E099–Transfer GPA	Continuous		
E101–Graduate studies flag	Dummy		
Financial Data			
X155–Financial need level	Categorical		
X122–Dependent/independent status	Categorical		
E126–Adjusted gross income	Continuous		
E370–Taxes paid	Continuous		
E373–Last amount collected	Continuous		
Regression Summary			
Pseudo R-square ^b		0.014	0.0166
Max rescaled R-square ^c		0.0402	0.0476
Df		10	16
Chi Square - likelihood ratio		331.0977	392.7721
Pr > ChiSq		<.0001	<.0001
PPC ^d		79.6	77.0

*p < 0.05.

**p < 0.01.

***p < .0001.

^a The variable was used in categorical form, grouping high school ranks into eight different levels (see Table 3). While the default rates of these groups were somewhat non-linear, overall higher ranks have lower default rates. Because this variable was used in a categorical form, the coefficient is positive rather than negative.

^{b,c} The pseudo R-square is a linear approximation of the percent variance explained by the model. It does not always extend over the full range of 0.0 to 1.0. The max rescaled R-square adjusts the pseudo R-square to the full range of 0.0 to 1.0 and thus is typically higher than the pseudo R-square. Both values are a rough approximation of the explanatory power of the model. “All the various R-square values...are low when compared to R-square values typically encountered in good linear regression models. Unfortunately, low R-square values in logistic regression are the norm and this presents a problem when reporting their values to an audience accustomed to seeing linear regression values.” (Holmes & Lemeshow, 2000.)

^d The percent predicted correctly was estimated as borrowers who had defaulted with a normalized model score of 0.60 or above, or those borrowers who had not defaulted with a normalized model score below 0.60. Model scores were normalized to a mean of 0.50 more closely to resemble the scores derived from the data mining process. Based on the data mining model, splitting the model scores at 0.60 rather than 0.50 reflected the maximum separation between defaulters and repayers in the data mining model.

Regression 3 Degree Info	Regression 4 GPA/Hours	Regression 5 Transfer/Grad	Regression 6 Financial	Regression 7 Data Mining
Transfer and Graduate Studies Data				
		-0.5026***	-0.5234***	
		0.0871**	0.1186***	
		0.2468	0.3309*	
Financial Data				
			1.5797	
			-8.4999	-29.6117***
			-0.00000395	
			-0.00012*	
			-0.00003	
Regression Summary				
0.0588	0.0676	0.0693	0.0704	0.0601
0.1682	0.1931	0.1981	0.2013	0.1719
22	27	30	35	7
1419.6442	1637.3353	1680.9696	1708.8295	1451.1909
<.0001	<.0001	<.0001	<.0001	<.0001
73.3	70.6	70.7	70.5	75.8