ROBERT H. SMITH SCHOOL OF BUSINESS, UNIVERSITY OF MARYLAND
BUDT 733 – DATA ANALYSIS FOR DECISION MAKERS

# Profiling Student Loan Deferment

December 12, 2007

UNIVERSITY OF MARYLAND
ROBERT H. SMITH SCHOOL OF BUSINESS
Leaders for the Digital Economy

***Prepared by:***
*Carlos E. Canales*
*April Humber*
*Irakli Nadareishvilli*
*Khalid Hasan*
*Soojung Jung*

## Executive Summary

Recent studies have shown that the cost of higher education has been rapidly increasing over the past few decades. The increase in costs is particularly more evident for students attending a university in the state of California, which has some of the best public academic institutions in the country. Quite often, many students attending California state school come from other states, thereby triggering Out-of State tuition costs which are usually 30% higher than the in-state tuition rate. In some instances, students seeking financial aid may receive some federal funding; however, federal funded grants and loans may not be enough to cover all academic expenses. Students in need of additional funds often have no choice but to seek private student loans.

In addition to qualifying for a loan, most lending institutions require students to make regular monthly payments while enrolled. The payments are low enough to be affordable for the student but often high enough to cover interest charges. In some instances, a student may face unforeseen circumstances, which require an academic leave of absence for an uncertain period. In these circumstances, federally insured loans can often be deferred for a grace period so long as the student is able to provide proof of hardship. On the other hand, deferment of private loans often involves more strict guidelines that could impede a student from deferring an existing loan, if necessary. Students experiencing unforeseen circumstances such as unemployment or a family death that requires a short academic leave of absence may wish to defer a loan and avoid default. Unfortunately, obtaining a private loan deferment can be quite difficult unless the student is well aware of the key characteristics lending institutions review prior to granting or denying a loan deferment.

This analysis has examined the set of characteristics used by private lending institutions to assess whether to grant or deny a loan deferment. The understanding of these characteristics is intended to provide valuable insight for a student wishing to increase their likelihood of obtaining a deferment on a private loan. The data source for this analysis is a sample database from the University of California, Irvine. It contains 1,000 observations, with 12 variables on student characteristics related to educational loans. Various data modeling tools such as logistic regression discriminate analysis, and classification trees were utilized to determine the most significant variables, which could be of interest for a student seeking loan deferment. The most significant variables characterizing student loan deferment are the *number of services* (e.g. Army, Air Force, and Navy) that a student belongs to, the *number of credit hours enrolled*, and the *number of months a student is absent*. All things considered, if a student wants to increase his/her likelihood of obtaining a loan deferment, they should join a service, enroll full-time, and/or limit their absenteeism.

## *Technical Summary*

Exhibit A contains a flowchart describing the process undertaken to determine the best model for characterizing student loan deferment. The details of the various steps are discussed in the proceeding sections.

### Data Exploration

Since the goal of our analysis was explanatory and the availability of continuous data was limited, we explored the data utilizing box plots. For variable selection purposes, we utilized classification trees to identify the most relevant predictors and also utilized pivot tables and statistical tests (chi-square test for independence). Exhibit B provides two insightful box plots of Deferred (the response variable) against *Number of Months Absent* and *Total Units Enrolled*. We see that those students whose loans were deferred had fewer months of absence and higher total units enrolled than those whose loans were not deferred. Based on domain knowledge, we were not surprised to discover that student loans are not likely to be deferred if a student is absent longer than 6 months on average. Also, students are more like to receive a loan deferment if they are enrolled, on average, in 9 or more units. From analyzing the classification tree, *Total Units Enrolled* was the first recursive split, indicating that it showed the most separation between classes and the next splits were on NUM_*MONTHS_ABSENT* and *TOTAL_UNITS*. We also utilized pivot tables and chi-square independence tests to analyze the correlation amongst variables. Through this process, we were able to narrow the down the nine original independent variables down to a set of six variables.

### Data Description

The following describes each of the original variables used to build the models on our analysis.

- Deferred – categorical response variable, binary with Y for deferred (not required to make loan payments while in school), N for not deferred.
- Sex – binary variable, F – Female student, M – Male Student
- Num_Months_Absent – continuous variable, number of months student has been absent
- Disabled – binary variable, Y- Disabled student, N- Not Disabled Student
- Bankruptcy – binary variable, Y- filed for bankruptcy, N – not filed for bankruptcy
- Employed – binary variable, Y – employed, N – unemployed
- Num _Units – continuous variable, number of units enrolled in 1st school
- Total_Units – continues variable, number of units enrolled in both schools
- Number _of _Services – categorical variable, student enlisted in 0, 1 or 2 services (armed or volunteer services (e.g fire department)

Based on the Classification Tree and graphs, several variables from the initial set were not included in the analysis. The excluded variables were School classifications, Service classifications, and Number of units enrolled for each school. Number of units enrolled in 1st school seems significant to loan deferment, however it was deleted because of high correlation to total number of units enrolled in both school. Students may be enlisted in 1 or 2 services. We created Number_of_services using the Service variables because a student enlisted in 0, 1 or 2 services may have some significance to student loan deferment.

## Explanatory Model Analysis

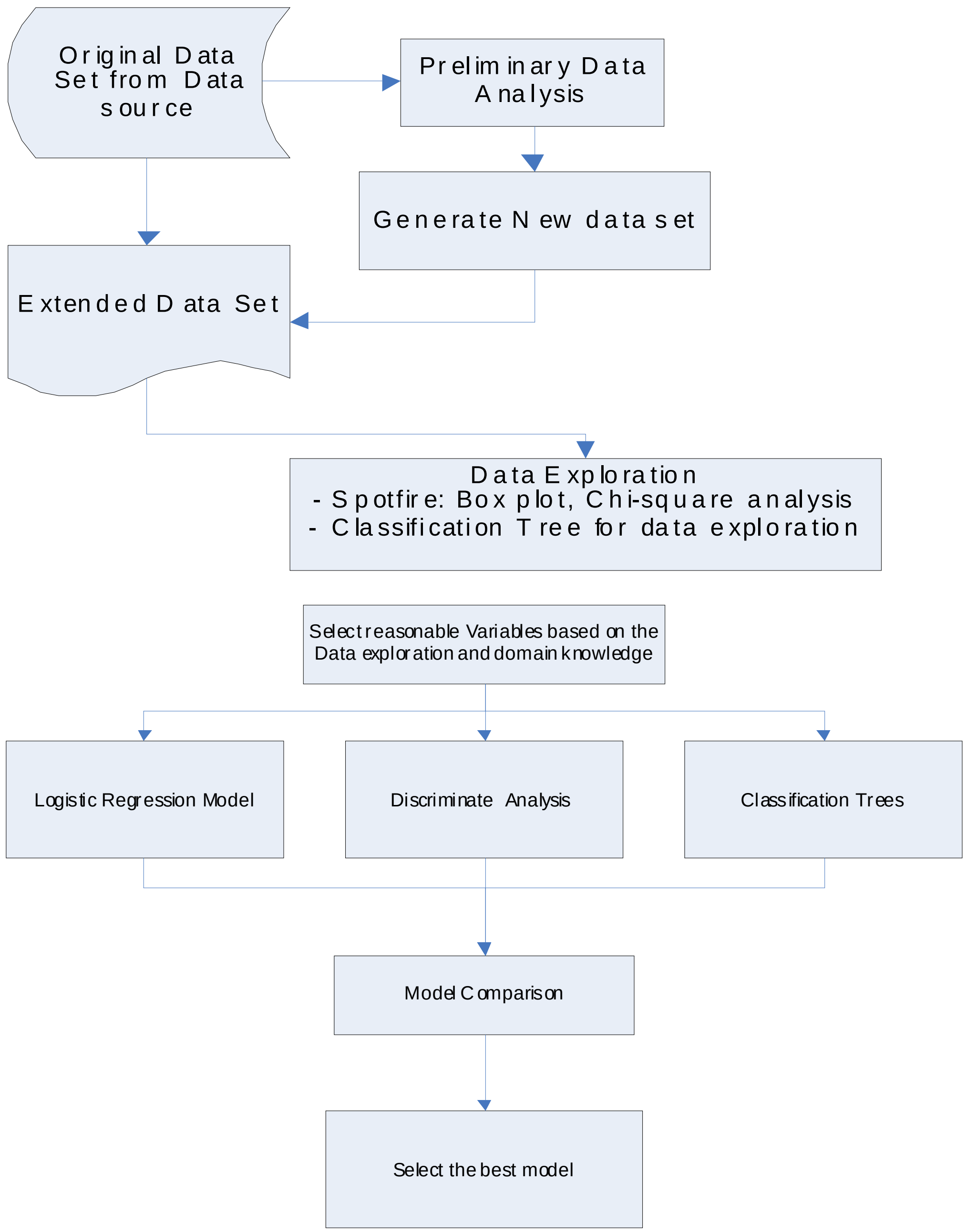
Based on domain knowledge, size of the data, and the explanatory goal, there were only three possible models that we could use in our analysis, Classification Trees (for data variable selection), Logistic Regression, and Discriminant Analysis using 'deferred' as the success class in each of the models. The classification tree indicated that Total Number of Units, Numbers of Month Absence, and Number of Services enlisted seem to be the more informative and useful.

We ran a Discriminant Analysis model using all the relevant variables, and noticed that the variables that best separated between deferred and not deferred were, in order of significance: Employment, Num of services, Total_Units, Num of months absent, Disabled, and Sex. It is interesting to note that Number_of_services, Total_Units, and Num_of_months_absent were also chosen as split variables in the classification tree. We next ran a Logistic Regression (LR) model, using the same relevant variables, and noticed a lower overall percentage error. Our final and best model is shown in Exhibit C, and has an overall error percentage of 19.9%. The R-Squared was 35.17% and the deviation was 844.92. We next obtained the deviance of the naïve model (1303.35) and tested whether the reduction from 1303.35 to 844.92 was statistically significant using Excel's CHIDIST function and got a 0 p-value. Thus, we concluded that our model is better than the naïve model from a statistical standpoint.

## Conclusion

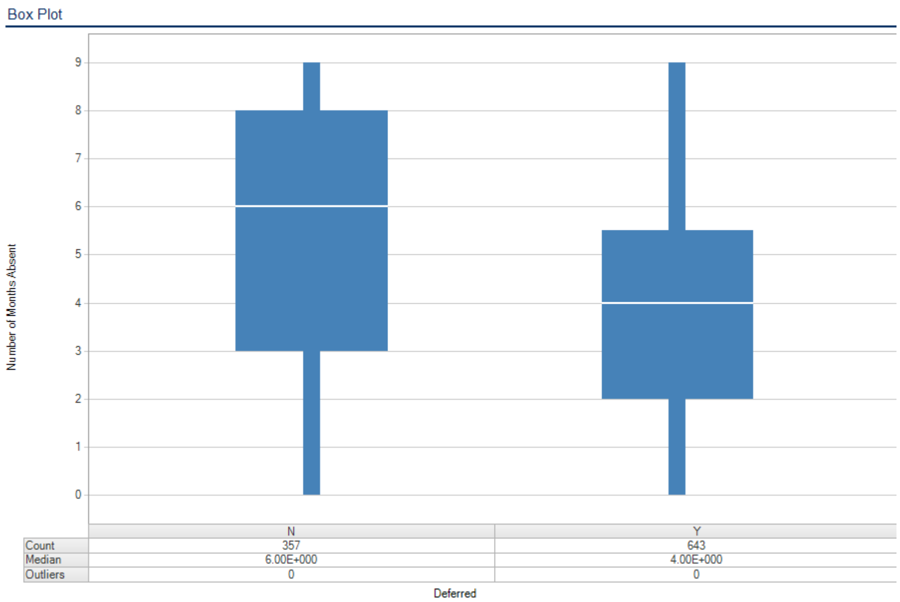
The logistic regression model (Exhibit C) is the best for characterizing student loan deferment because it is reasonable, parsimonious, and offers the easiest explanation. The LR model highlights that the *Number of Services*, *Total Units*, and *Number of Months Absent* best explain differences in whether a student's loan is deferred or not. In particular, holding all other variables constant, each: (1) additional service that a student belongs to is associated with a 4.12 increase in the odds of his/her loan being deferred, (2) additional unit (credit hour) that a student takes is associated with an increase in the odds of obtaining a loan deferment of 1.40, and (3) additional month that the student is absent from school, decreases the odds of receiving a loan deferment by 0.73. The logistic regression model also highlights that sex, disability status, and employment status do not do a good job of discriminating student loan deferment. In general, we have reverse engineered the process used by commercial banks in California to grant deferments to student loans and determined that if students want to increase the likelihood of deferring their loans, they should take a full-time course load, enroll in a service (Army, Air Force, etc), and/or limit their absenteeism.

# Exhibit A – Modeling Process Flowchart

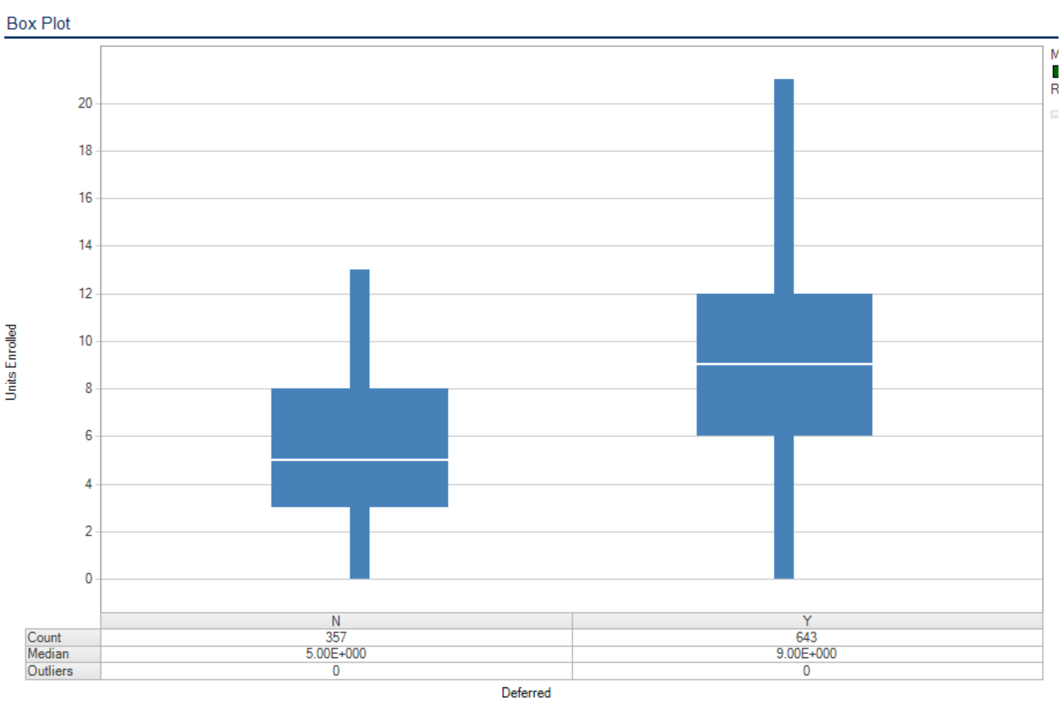


# Exhibit B – Box Plots

Box Plot of Deferred versus Number of Months Absent




Box Plot of Deferred versus total Number of Units Enrolled



## Exhibit C – Final Logistic Regression Model

| Input variables | Coefficient | Std. Error | p-value | Odds |
|---|---|---|---|---|
| Constant term | 16.84710503 | 540.8853149 | 0.97515208 | * |
| NUM_MONTHS_ABSENT | -0.31529847 | 0.0337925 | 0 | 0.7295711 |
| DISABLED | 0.39086822 | 0.35954201 | 0.27698022 | 1.47826374 |
| Sex | 0.13317721 | 0.17069066 | 0.43525809 | 1.14245248 |
| Employed | -17.84041405 | 540.8852539 | 0.97368759 | 0.00000002 |
| TOTAL_UNITS | 0.33731708 | 0.02608318 | 0 | 1.40118325 |
| Number of Services | 1.41654301 | 0.20000213 | 0 | 4.12284327 |

| | |
|---|---|
| Residual df | 993 |
| Residual Dev. | 844.9161377 |
| % Success in training data | 64.3 |
| # Iterations used | 20 |
| Multiple R-squared | 0.35173258 |

| Classification Confusion Matrix | | |
|---|---|---|
| | Predicted Class | |
| Actual Class | Y | N |
| Y | 558 | 85 |
| N | 114 | 243 |

| Error Report | | | |
|---|---|---|---|
| Class | # Cases | # Errors | % Error |
| Y | 643 | 85 | 13.22 |
| N | 357 | 114 | 31.93 |
| Overall | 1000 | 199 | 19.90 |