USING DATA MINING TEHNIQUES TO IDENTIFY "THE BEST" OPERATIONAL PATTERNS FOR ENROLLMENT MODELING

A THESIS

SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

FOR THE DEGREE OF MASTER OF SCIENCE

IN THE GRADUATE SCHOOL OF THE

TEXAS WOMAN'S UNIVERSITY

COLLEGE OF ARTS AND SCIENCES

BY

BOGDAN CATALIN OBARSE, B.A.

DENTON, TEXAS

AUGUST 2009

TEXAS WOMAN'S UNIVERSITY
DENTON, TEXAS

May 11, 2009

To the Dean of the Graduate School:

I am submitting herewith a thesis written by Bogdan Catalin Obarse entitled "Using Data Mining Techniques to Identify "The Best" Operational Patterns for Enrollment Modeling". I have examined this thesis for form and content and recommend that it be accepted in partial fulfillment of the requirements of the degree of Master of Science with a major in Mathematics.

Mark S. Hamner

Mark Hamner, Ph.D., Major Professor

We have read this thesis and recommend its acceptance:

David Marshall, Ph.D

Ellina Grigorieva, Ph.D

Don E. Edwards

Don Edwards, Ph.D
Department Chair

Accepted:

Jennifer Martin

Dean of the Graduate School

ACKNOWLEDGEMENTS

I will like first to thank Dr. Hamner for his support and invaluable guidance. Without his patience and encouragement I would not been able to complete this thesis. I would also like to thank Dr. Edwards and Robyn for all their help before and after I was able to come here. Finally, thanks to my parents and my friends, for their support throughout this undertaking.

ABSTRACT

BOGDAN CATALIN OBARSE

USING DATA MINING TECHNIQUES TO IDENTIFY "THE BEST" OPERATIONAL PATTERNS FOR ENROLLMENT MODELING

AUGUST 2009

For any Educational Institution it is very important to know the number of new students and the number of returning students. Based on these numbers, there could be conducted predictions of the budget that the institution will have for the next year.

This research will utilize pre-existing historical data from Texas Woman's University containing readily available and easily measured factors, which most institutions of higher learning will have available, and will split the existing data in all the sub sets possible. Running a chi square analysis on each set obtained, the program will be able to show us which splitting way is better for obtaining the most consistent patterns, using the provided data. The results will be compared with the results obtained running a linear regression analysis on the same data sets.

The study will introduce an extraneous hidden-time variable related to partitioning ways possible.

The program can be used in the future on any University data sets, providing the most holding combination of variables that will hold over the years.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER I

INTRODUCTION

The notion of a viable model to predict a dichotomous variable such as whether a student will enroll or not enroll at a specific university entails either an explicit or implicit fundamental supposition that there are underlying patterns that will be useful and reliable for accurate forecasting of the variable of interest. Yet, any theoretical model that explains the variability or pattern that underlines the variable of interest is dependent on observable data a researcher has available. In general, the patterns that the model captures will provide accurate predictions if these patterns are sustained in future data related to the variable of interest. In other words, fitting a model on observed or historical data that has underlying patterns to explain the variable of interest will be useful for accurate prediction, if those underlying patterns in future data mimic or hold to the patterns from which the model was built. For many researchers, available data needed for modeling purposes comes from snap shots of data coming from the time of interest. For example, institutions of higher education studying the dichotomous variable of retention will have snap shots of fall census data, spring census data, and summer census data from which to explore and use to make reports and/or for modeling. This study will create a data mining program that is used as a pre-modeling technique to explore a multitude of partitioning and stratification combinations on snap shots of observed historical data to determine which combination generates the "best" sustained patterns for the dichotomous variable of interest. In addition, we will derive and specify the underlying statistical decision rule that determines the "best" partitioning patterns within our data mining program. Thus, this study will provide a pre-modeling tool that determines the most reliable or best, as defined by our decision rule, partitioning patterns that can be used for predictive modeling.

In this research we will illustrate a pre-modeling technique on snap shots of data by utilizing pre-existing historical data from Texas Woman's University (TWU) containing readily available and easily measured factors or variables, which most institutions of higher learning will also have available. To implement this pre-modeling technique we will write a computer program that will efficiently explore the variable of interest contained in the snapshots of historical data. Through our derived decision rule, the program will iterate through a multitude of partitions and stratifications of these combined snapshots of historical data to determine the partition and stratification method that yields the most consistent or "best" historical patterns for predictive modeling. Thus, regardless of the model used for prediction, we will illustrate using TWU data that attention to partitioning and stratification patterns of the data can enhance prediction.

In Chapter-II, we will review the literature on predictive modeling concerning higher education data. Chapter-III will introduce the idea of partioning snap shots of historical data and provide a simple example to illustrate that partitioning snap shots of data can alter the historical patterns on the dichotomous variable of interest. Chapter-III will also introduce appropriate definitions and notation needed to derive our decision rule. In Chapter-IV, we will define and illustrate how the decision rule works. Using this decision rule our program will provide the best historical enrollment patterns found from TWU snap shots of data. Chapter-V will model the top three historical patterns determined by our program and compare their predictive accuracy on data that was not used to make the model. Finally, Chapter-VI will provide our conclusions and a discussion of possible future research.

CHAPTER II

LITERATURE REVIEW

Many researchers in higher education have an interest in understanding their student population. To facilitate their understanding, institutions will likely explore the data they collect to determine viable patterns that can lend to more insights to the students they serve. In this regard, the institution may be interested in predictive modeling, especially when it comes to predicting enrollment since funding for public universities may depend on the total credit hour production their enrolled students generate. To predict re-enrollment at an institution of higher education requires two fundamental components: student information data from the past and a mathematical model for prediction. Regardless of the issue that may concern those involved with an institution of higher education, in general, an interest in using data to facilitate understanding of their student clientele through reporting or to generate predictive modeling is common. In this chapter, we will review various uses of student information data and note that the literature in higher education lacks a data mining technique that explicitly uses a statistical decision rule to determine historical consistency of snap shots of data.

Any institutions geographical location, employees, size, and services certainly contribute to the general culture of that institution and help to define its unique atmosphere. Thus, it stands to reason that variable selection and data gathering needed to understand an institutions clientele might vary across institutions. In this regard, a paper by the consulting group Noel Levitz (2002) addressed this issue making a note that, while no two institutions are ever alike in the exact data, they hold many data needs in common. Noel Levitz makes the point that by tracking these data and derivatives that apply to each situation, they have literally seen more than Max(P) enrollment managers making changes necessary to sustain a competitive edge. In a study by Goenner and Snaith (2004), they note that theory should dictate this choice of variables in modeling. However, they acknowledge that often in social science there are several theories that may suggest the inclusion or exclusion of certain variables as controls. The result of this is that researchers may use different variables in their analysis and come to different results with respect to predicted effects and their statistical significance. Nonetheless, Goenner and Snaith (2004), note that empirical analysis requires researchers to choose which variables to use as control in their models. The implication is that having necessary data to determine significant variables can facilitate understanding of their clientele and such data may be useful in modeling so that an institution can be competitive in their management of enrollment.

For an institution to understand their student clientele, variables and the accompanying data are needed so that empirical analysis and reporting can be used to facilitate administrative decisions regarding the clientele the institution typically attracts. Before beginning to build a model or gain information from data, it is important to perform an exploratory data analysis (EDA). EDA is a set of procedures aimed at understanding the data and the relationships among the variables (Refaat 2006; Hoaglin, Mosteller, Tukey and John Wilder 1985; Valleman, Hoaglin 1977). In this regard, to identify appropriate variables and patterns that may be useful or considered significant, from a statistical standpoint, generally involves programming skills to mine the data or statistical modeling skills to use the data for predictive modeling. For example, William N. Anderson (2005), makes a deliberate point to mention that before conducting a research study, people involved must first seek assistance from colleagues from their institutional research officer or from the faculty members in mathematics, statistics or operations research department to assure those involved have the right skill set to address the problem. A study made by Goenner and Pauls (2006) found that statistically significant variables such as geographic and demographic data based on the student's zip code provide predictors for enrollment. Nonetheless, all the aforementioned studies do not address a variable's significance in relation to how the data is partitioned.

When it comes to prediction, there are many theoretical models that an institution can use in order to aid their prediction. The types of model that can be used are predicated on the variable of interest. For example, William N. Anderson (2002) used Linear Regression to Predict the Number of Students in a Freshman Class. In general, a linear regression model is appropriate, assuming appropriate assumptions have been met, to model a qualitative variable. Another example of linear regression involved modeling undergraduate graduation rates (Hamrick 1964). However, when it comes to the enrollment of an individual, the variable is qualitative (i.e. categorical) with a dichotomous outcome such as enroll or not enroll. For a dichotomous variable, logistic regression is often used as a model. For instance, enrollment prediction is quite frequently modeled utilizing logistic regression (Ahluwalia 2006; Goenner and Pauls 2006; Morley 2000). Predicting retention in institutes of higher education is another example of where logistic regression was used (Astin 1997; Hamrick, Schuh and Shelley 2004; Hurtado, Saenz, Santos and Cabrera 2007). Yet another example is predicting success in college where the outcome is completion versus non-completion of a bachelor's degree (Geiser and Santelices 2007; Trusty and Niles 2003). In addition, logistic regression has been used to model success in particular courses (Belcheir 2002; Perkhounkova, Noble, and Sawyer 2005). In general, those models specify the parameter estimates obtained from using significant variables from modeling observed data, but their data are not made by combining snap shots.

In the literature there are qualitative variables of interest that may not be dichotomous in nature. Many researchers may be interested in predicting the outcome of a variable with three or more categories, known as *polychotomous*. For instance, the variable 'type of service use' is an outcome variable that is often measured in health services research. Types of health services utilization can include medical provider visit, hospital outpatient visit, emergency room visit, hospital inpatient stay, and home health care visit (Hedeker 2003). Researchers in the medical sciences also make extensive use of polychotomous logistic regression modeling when assessing risk of disease (Dubin and Pasternack 1986; Lawrence et al. 2006; O'Shea et al. 1999). Social science is yet another field of research in which this statistical analysis is performed. For instance, Koivusilta, Arja, and Andres (2003) utilized polychotomous logistic regression analysis to assess the associations between health behaviors and health in adolescence and attained educational level in adulthood. Another example from the social sciences assesses the relationship between perceived life satisfaction and substance abuse in adolescents (Zullig et al. 2001). Similar to the dichotomous modeling techniques mentioned above, these polychotomous modeling techniques were applied to a certain subset of data and did not address expanding the discussion to a combination of snapshots of data.

For any study which specifies significant variables or indicators, the reliability of the variables usefulness as a predictor or as a viable indicator to help administrators understand student enrollment will be validated if the pattern from which its significance was determined generally holds for future data. In the context of modeling for prediction, this means that parameter estimates for the predictive model, from the historical data, will generally imitate the patterns or parameter estimates you would expect to get with the current data after the results of reenrollment are actually known. This particular nuance with the historical patterns towards variable patterns for prediction or reporting is rarely mentioned in the literature. In addition, none of the studies we have seen in the literature explicitly address the importance of partitioning their data sets prior to model building in order to facilitate better patterns. Rather they typically specify parameter estimates from modeling a single set of data.

In one study, Marshall and Oliver (1979) proposed a forecasting model, on a set of current students, based on the reenrollment patterns of certain historical first-time students. These sets of first time students were found by finding the historical semester data from which the students of interest first entered the University. The unmentioned importance of modeling the enrollment patterns of first-time students is that you implicitly create mutually exclusive partitioning of the historical snap shots of data so that a student of interest will only belong to a single historical set, which will help avoid over prediction. No less important in Marshall and Oliver's paper is the unaddressed discussion about modeling different mutually exclusive partitions of the historical data to possibly yield better predictions.

In this study we will develop a data mining technique to uncover relevant patterns in the snap shots of data prior to model building. After the academic data from students has been cleaned up and in a proper format through the use of data mining, the exploratory data analysis conducted in this study will provide a method to determine a partitioning of the data which provide historically viable holding patterns for modeling the dichotomous outcome of interest. In the next chapter we will introduce the notation necessary to help explain how you can partition data into various mutually exclusive sets and show, through an example, that partitioning the historical snap shots of data does make a difference in the observed patterns.

CHAPTER III

DATA EXPLORATION

To gain understanding of a certain topic of interest, a conceptualization of the problem itself and the factors or variables that can influence the phenomenon is essential. From conceptualizing the problem and identifying the appropriate factors of interest, administrators or many businesses these days have the capability to store and retrieve data concerning these factors to better understand the clientele they serve or intend to serve. In a predictive modeling paper by Thomas, Dawes and Reznik (2001), they make the point that an institution needs to have "good data in a usable form." Having good data can involve cleaning data entry errors and also keeping appropriate variables or fields within the database that can help understand the topic of interest. In predictive modeling or for any kind of reporting, a conceptualized variable or factor is not usable if the field within the database, which stores its values, is not well occupied. Nonetheless, good data in and of itself is not inherently informative, it still has to be explored or mined appropriately in order to determine usable or reliable patterns that will facilitate modeling or even simple reporting. For example, Marshall and Oliver (1979, p. 196) modeled the "very stable behavior in student attendance." Implicit in this statement is that historical patterns were explored and the patterns that exhibited consistency, presumably over time, were deemed worthy for use in predictive modeling. In our study we will explore patterns of data that is formed by composite snap shots. Composite snap shots of data, as defined below, is a combination of snap shots of data frozen at specific moments of time, In this chapter we will introduce necessary mathematical notation that will facilitate our discussion and to help illustrate that partitioning composite snap shots of data into mutually exclusive sets is an important consideration (i.e. matters) when it comes to finding stable or reliable patterns over time that can be used for reporting or modeling.

Definition 1. Let S_t represent a snapshot of data at some time $T = t$. Without loss of generality suppose $T = 1, 2, \ldots, l$ corresponds to points of time with corresponding snapshots of data $S_1, S_2, \ldots, S_l$. Then a *composite snapshot of data* is the set $P = S_1 \cup S_2 \cup \cdots \cup S_l = \bigcup_{i=1}^{l} S_i$.

The idea of composite snap shots is very common with higher education data and can be found in many other situations. Our particular example will illustrate how we make a composite data sets using snap shots of semester data and explore that data in order to understanding reenrollment patterns. Thus, we will explore patterns from the data related to the dichotomous state of whether a set of students reenrolls or does not reenroll into a semester of interest. Using notation we will let $P = \{1, 2, \ldots, N\}$ represent the set of indices for the finite population of potential reenrolling students of size N. The set P represents the total number of unique individuals we expect from our composite data sets. For each individual $k \in P$, the variable of interest or the dependent variable is defined as

$$y_k = \begin{cases} 1 \text{ , Student reenrolls at TWU} \\ 0 \text{ , Student does not reenroll at TWU} \end{cases} \tag{3.1}$$

Suppose the administration wants to understand the patterns of y_k for the current academic year of students to help determine who will return the following academic's year Fall semester. An *academic year* for an institution of higher education begins in the Fall semester, continues into the Spring Semester, and ends after the summer semester prior to the following fall semester, which begins a new academic year. For example, the current academic year consists of the following semesters: Fall 2008, Spring 2009, and Summer 2009. It's worth noting that because the Fall semester and Spring semester, for example, of the same academic year occur at different calendar years, we will not use a calendar year to refer to an academic year in this paper. We will assume it is understood that when we refer to, for example, the Fall Semester and Spring semester from the same academic year that the reader realizes those semesters do not occur under the same calendar year. In our example, we consider students that attend the fall or spring of the same academic year. The composite of those snap shots of data give us a set of individuals, where the individuals are listed only once. Figure 1 illustrates the idea of a composite data set formed by the combination of two snap shots of data from the same academic year.

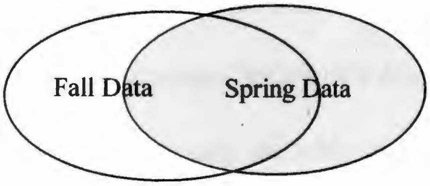

Figure 1. Composite Set.

Using mathematical notation, we will define how to obtain a composite data set from the snap shots of data represented by Figure 1. Suppose that you have data from the Fall and Spring semesters of the same academic year labeled as FA and SP respectively. Using set notation to facilitate this discussion, our composite data set is defined as $P = FA \cup SP$, where P is a set of size N consisting of the set of indices for students coming from the union of FA and SP data sets. It is worth noting that the way we defined P as the union of multiple sets, from a mathematical stand point, each unique student receives a single index. This means that although there may be students who happen to be in the intersection of the fall and spring snap shots of data (i.e. students that were enrolled in the fall as well as the spring semesters) only a single index is assigned to each of them.

Having one index for each unique student in the set P is particularly important for predicting. For example, suppose the researcher wants to predict if those students from the current academic year FA and SP semester will reenroll the next Fall semester, which is the beginning of the following academic year. A mistake the researcher could make is to predict the reenrollment of FA students and then predict the reenrollment of SP students, since many students will be listed in both data sets. In other words, an intersection exists between the sets FA and SP which means FA and SP are not *mutually exclusive* since $FA \cap SP \neq \varnothing$. Consequently, over prediction will be a problem. Thus, the researcher should consider making a prediction on sets of students that are mutually exclusive and together form a *partition*, see definition below, of the set P.

Definition 2. Let $k \in A_i$, a *partition* of the set P consists of sets $A_1, A_2, \ldots, A_k$ such that P $= A_1 \cup A_2 \cup \cdots \cup A_k = \bigcup_{i=1}^{k} A_i$ where $A_i \cap A_j = \varnothing$ for $i \neq j$.

From Definition 1, the composite of l snap shots of data P = $S_1 \cup S_2 \cup \cdots \cup S_l$ is composed of sets for which an intercept is possible. In other words, for any two sets S_i and S_j for which $i \neq j$ the possibility exists that $S_i \cap S_j \neq \varnothing$, for any three sets S_i, S_j, S_k for which $i \neq j \neq k$, the possibility exists that $S_i \cap S_j \cap S_k \neq \varnothing$, and for four sets up to l

sets a similar argument of possible intercepts existing can be made. So, given a composite of l snap shots of data P = $S_1 \cup S_2 \cup \cdots \cup S_l$ and the knowledge that possible intersections exists, we can now count the maximum number of mutually exclusive sets that can be formed from S_1, $S_2, \ldots, S_l$. To illustrate our discussion on counting these sets we will periodically refer to Figure 3, which contains only $l = 3$ snap shots of data: $S_1 = A_1 \cup A_4 \cup A_5 \cup A_7$; $S_2 = A_2 \cup A_5 \cup A_6 \cup A_7$; $S_3 = A_3 \cup A_4 \cup A_6 \cup A_7$. First, for l sets of snap shots, the maximum number of subsets that do not contain an intersection is

$$\binom{l}{1} = \frac{l!}{(l-1)!\,1!},$$

which we will denote as ${}_lC_1$. For example, in Figure 3, where $l = 3$, we can see that the maximum number of subsets that individually are not composed of an intersection is ${}_3C_1 = 3$, these subsets are A_1, A_2, and A_3. Again, for l sets of snap shots, the maximum number of only pair wise intersections is

$$\binom{l}{2} = \frac{l!}{(l-2)!\,2!},$$

which we will denote as ${}_lC_2$. For example, in Figure 3, where $l = 3$, we can see that the maximum number of only pair wise subsets is ${}_3C_2 = 3$. These subsets are A_4, A_5, and A_6.

Next, for l sets of snap shots, the maximum number of subsets that are composed by the intersections of exactly three sets is

$$\binom{l}{3} = \frac{l!}{(l-3)!\,3!},$$

which we will denote as ${}_lC_3$. For example, in Figure 3 we can see that the maximum number of subsets that are composed by the intersections of exactly three sets is ${}_3C_3 = 1$. This is the subset A_7. Following this pattern for the general case when we have l snap shots of data that compose P $= S_1 \cup S_2 \cup \cdots \cup$ S$_l$, the maximum number of mutually exclusive subsets, created from considering all possible intersections is

$$\text{Max}(P) = \sum_{i=1}^{l} \binom{l}{i} = 2^l - 1 \qquad (3.2)$$

An important observation of these Max(P) subsets, is that these subsets are mutually exclusive and together form a partition of P. For example, in Figure 2, we see that $A_i \cap A_j = \varnothing$, for $i \neq j$, and $P = A_1 \cup A_2 \cup A_3 \cup A_4 \cup A_5 \cup A_6 \cup A_7$. In general, for a composite of l snap shots of data $P = S_1 \cup S_2 \cup \cdots \cup$ S$_l$, we can rewrite P as partition of Max(P) mutually exclusive sets

$$P = A_1 \cup A_2 \cup \cdots \cup A_{\text{max(P)}}. \qquad (3.3)$$

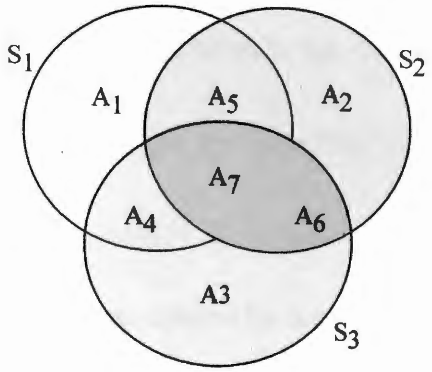

Figure 2. Mutually Exclusive sets

Equation 3.3 represents a partition of $P = S_1 \cup S_2 \cup \cdots \cup S_l$ that would contain the maximum number of subsets. However, there are many ways to partition a composite data set P that will have $k <$ Max(P) subsets that are mutually exclusive. In general, an alternative partition of the set P is constructed by considering the possible associations of the mutually exclusive subsets A_i, $i = 1, 2, \ldots,$ Max(P). To make a Max(P)-1 partition we can simply associate any two subsets A_i and A_j, $i \neq j$, for example, $\mathrm{P} = (\mathrm{A}_1 \cup \mathrm{A2}) \cup \mathrm{A}_3 \cdots \cup \mathrm{A}_{\mathrm{max(P)}} = A_1^* \cup A_2^* \cup \cdots \cup A^*_{Max(P)}$, where $A_1^* = (\mathrm{A}_1 \cup \mathrm{A}_2)$, $A_2^* = \mathrm{A}_3, \cdots, A^*_{Max(P)-1} = \mathrm{A}_{\mathrm{max(P)}}$. Notice that $\mathrm{P} = A_1^* \cup A_2^* \cup \cdots \cup A^*_{Max(P)-1}$ satisfies Definition 2 since $A_i^* \cap A_j^* = \varnothing$ for $i \neq j$. Similarly, we can create combinations of associations, A_i^*, created from two or more subsets of $A_1, A_2, \ldots A_{max(P)}$, to create partitions of P, which can contain anywhere from 1 to $Max(P) - 1$ subsets. However, in this paper, we generally restrict an association to one of the following conditions: i) $A_i^* \subseteq S_j$ or ii) For any sequence length $k \leq \max(P)$ of consecutive snap shots S_t, $S_{t+1}, \ldots, S_{t+k}$, $A_i^* = S_t \cup S_{t+1} \cup \cdots \cup S_{t+2}$ only if $S_t \cap S_{t+1} \cap \cdots \cap S_{t+k} \neq \varnothing$. The one exception to condition (i) and (ii) is the partition created by the following association: $P = (A_1 \cup A_2 \cup A_3 \cdots \cup A_{\max(P)}) = A_1^*$.

In Figure 2 we illustrate two methods of partitioning $P = FA \cup SP$ using the following mutually exclusive subsets: 1) *Fall-Main* = *FA* and *Spring*_Subset = *SP* – *FA* (or $SP \cap (FA \cap SP)'$) ; 2) *Fall*_Subset = *FA* – *SP* (or $FA \cap (FA \cap SP)'$) and *Spring-Main* = *SP*. Method 1 creates the following partition P = *Fall-Main* ∪ *Spring*_Subset and Method 2 creates the partition P = *Fall*_Subset ∪ *Spring-Main*. Notice that P = FA ∪ SP satisfies the definition of a composite data set, Definition 1, but does not satisfy the Definition 2 of a partitioned data set until we rewrite P using either Method 1 or Method 2 above.

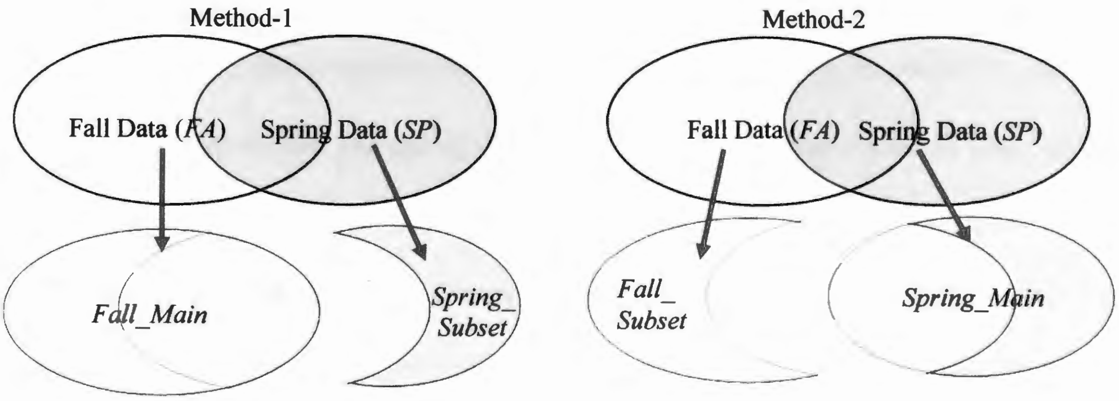

Figure 3. Basic Partitioning of Academic Year Data

Based on basic methods of partitioning from Figure 2, we will illustrate using a simple example that partitioning can play an important role in finding consistent patterns of reenrollment. Table 1 illustrates patterns of enrollment for two years using several methods of partitioning for each year. To illustrate our earlier notation, let P_1 represent the set of indices for the finite population of potential reenrolling students from academic Year 1of size $N_1 = 255$ and let P_2 represent the set of indices for the finite population of potential reenrolling students from academic Year 2 of size $N_2 = 254$. It is worth noting that in our actual data exploration we found that most students in the fall semester will also reenroll the following Spring semester. Hence the intersection, which we now denote as $Intersection = FA \cap SP$, see Figure 2, will be a larger subset of P than Fall_Subset or Spring_Subset. Although this example is contrived and does not illustrate actual observed data, we see in Table 1 under Method 3 that the intersection for both Year 1 and Year 2 data constitute the largest subsets of P_1 and P_2 respectively. This phenomenon that the fall and spring data, coming from the same academic year, we believe, is common for most institutions of higher education. Also, notice that in Table 1, for the respective years, the total number of students, the total number of returning (y=1) students, and the total number of students that do not return (y = 0) is exactly the same for each method of partitioning. To better see the patterns of reenrollment given by the different partitioning methods P_1 and P_2 in Table 1, we will present a table that shows the reenrollment patterns of Table 1 in terms of percentages.

Table 1

Re-enrollment Patterns of $P = FA \cup SP$ by Partition

		Year-1			Year-2		
Method	Partition	No. Students	Return (y = 1)	Not Return (y = 0)	No. Students	Return (y = 1)	Not Return (y = 0)
Method 1	Fall_Main	175	135	40	175	134	41
	Spring_subset	80	70	10	79	69	10
Total (N)		255	205	50	254	203	51
Method 2	Fall_subset	75	60	15	75	64	11
	Spring_Main	180	145	35	179	139	40
Total (N)		255	205	50	254	203	51
Method 3	Fall_subset	75	60	15	75	64	11
	Intersection	100	75	25	100	70	30
	Spring_subset	80	70	10	79	69	10
Total (N)		255	205	50	254	203	51

Recall that our objective is to explore historical data in order to find patterns that exhibited consistency over time. In Table 2 we see an example of two years worth of historical reenrollment patterns presented by different partitioning methods in terms of percentages. In this context Table 2 shows that for Year 1 under partition Method 1 that 77.1% of the 175 students in Fall_Main returned while 87.5% of the 80 students in spring subset returned. Notice that in Year 2 similar reenrollment percentage patterns emerge: 76.1% of the Fall_Main returned while 87.3% of the Spring_Subset returned. From Year 1 to Year 2 the reenrollment percent patterns in Fall_Main were off by only .5%.

Similarly, the reenrollment percent patterns in Spring_Subset were off by only 0.2% from Year 1 to Year 2. Analyzing the data from this perspective, we see the most consistent enrollment patterns from Year 1 to Year 2 are found under partition Method 1 which shows less variation in enrollment patterns from year to year in terms of percentages. From Table 2 we see that the cumulative absolute value of the differences in enrollment percentages by partition Method 1 from Year 1 to Year 2 is only 1.4%. Partition Method 2 is a distant second in terms of enrollment consistency by having a cumulative absolute value of the differences in enrollment percentages from Year 1 to Year 2 of 16.4%. This example illustrates that partitioning matters when trying to find consistent historical enrollment patterns.

Table 2

The Percent of Enroll and Not Enroll by Partitioning Method

		% Return (y = 1)			% Not Return (y = 0)		
Method	Partition	Year 1	Year 2	% off	Year 1	Year 2	% off
Method 1	Fall_Main	77.10%	76.60%	0.50%	22.90%	23.40%	-0.50%
	Spring_subset	87.50%	87.30%	0.20%	12.50%	12.70%	-0.20%
Total off (Absolute Value)				0.70%			0.70%
Method 2	Fall_subset	80%	85.30%	-5.30%	20%	14.70%	5.30%
	Spring_Main	80.60%	77.70%	2.90%	19.40%	22.30%	-2.90%
Total off (Absolute Value)				8.20%			8.20%
Method 3	Fall_subset	80%	85.30%	-5.30%	20%	14.70%	5.30%
	Intersection	75%	70%	5%	25%	30%	-5%
	Spring_subset	87.50%	87.30%	0.20%	12.50%	12.70%	-0.20%
Total off (Absolute Value)				10.50%			10.50%

For institutions of higher education there is typically useful information that is known concerning their students. For now we have introduced the idea of partitioning a set P that adheres to Definition 2 in order to find consistent patterns for a single dependent variable, y_k, as defined by Equation 3.1. Yet, for each individual $k \in P$ a set of p independent variables

$$\mathbf{x}_k = \left\{x_{1k}, x_{2k,} \ldots x_{pk}\right\}, \tag{3.4}$$

will exist that can facilitate understanding of the dependent variable, y_k. These independent variables can also facilitate partitioning and bring about an interesting consideration a researcher has to address when making certain types of partitioning on snap shots of data. Recall snap shots of data make up the set P. For example, we introduced the set P = FA $\cup$ SP. These snap shots of data are made at different times and hence the information contained in $\mathbf{x}_k$ for each k $\in$ P can change. So, for a set such as *Intercept* = *FA* $\cap$ *SP*, should we just use the latest information in $\mathbf{x}_k$? For example, a student in FA may be classified as a sophomore in FA, but in the same academic year and subsequent semester SP they may be classified as a junior. In other words, should we tag or assign to the individuals in *FA* $\cap$ *SP*, the information $\mathbf{x}_k$, contained in FA or SP? For categorical variables such as classification, the answer depends on whether stratification using that variable on the set we named *Intercept* facilitates finding more consistent patterns of y_k using the information from $\mathbf{x}_k$ from either snap shot FA or snap shot SP.

As mentioned previously, good data in and of itself is not inherently informative. It still has to be explored or mined appropriately in order to determine usable or reliable patterns that will facilitate modeling or even simple reporting. Thus, our discussion concerning partitioning and stratification of data in order to find consistent patterns of y_k, has introduced the importance of a data mining process that constitutes a "pre-modeling" technique. This type of discussion is not explicitly found in the literature. Yet, this pre-modeling technique will allow us the ability to know which sets A_i, $i = 1, ..., k$, as defined by Definition 2, form a partition of P for which, as Marshall and Oliver (1979, p. 196) state, provide "very stable behavior in student attendance" that can be modeled in order to obtain a good prediction.

To implement this pre-modeling technique will require considering many combinations of sets that form a partition of the set P. To efficiently consider many partitions of the set P, we will develop a data mining algorithm that can be applied to any composite data set composed of snap shots of data in order to identify significant patterns. In Chapter IV we will specify the statistical decision rule that will determine the most consistent patterns of y_k by selecting the appropriate subsets A_i, $i= 1, ..., k$, that form a partition of P.

CHAPTER IV

DECISION RULE

In Chapter III we discussed the different ways that we can partition the data. In fact, the more snapshots of data that defines the composite data set P, then the more ways the set P can be partitioned. Having $(Max(P) \times (k))$, where k is the number of levels in a categorical variable, there are over Max(P) of possible combinations to make a partition. The need for a computer algorithm becomes imperative in order to explore these possibilities. In this chapter we discuss how our computer algorithm will explore various partitioning methods and specify the decision rule we will create in order to determine the "best" partition that shows consistent historical patterns on y_k.

We are interested in knowing if there are significant reenrollments patterns from students attending TWU during the Fall and Spring semester of the same academic year. In particular we would like to facilitate finding reenrollment patterns on the set of students indexed by the set P = FA ∪ SP by determining an appropriate partition of P geared towards finding consistent historical enrollment patterns. Thus, our computer algorithm will be limited to the historical composite data sets P = FA ∪ SP, but our discussion can easily be extended to include more snap shots of data that form a composite. Although we are using only two snapshots of data to form our composite data set, it is worth noting that exploratory analysis on TWU historical data found that for any fall semester of interest, a majority of the undergraduate reenrolling students come from the previous academic year fall and spring semesters: $P = \text{FA} \cup \text{SP}$. For example, in Table 3, we see that for fall 2001 through fall 2003 the proportion of total re-enrollment coming from the set P is consistently around 90%. For graduate students the proportion of total re-enrollment coming from P is generally over 65%. The phenomenon that a majority of reenrolling students come from P would likely be the situation for most institutions of higher education

Table 3

Undergraduate Reenrollment

			Fall Total			Enroll Percent (%)		
Status	Student Level	Reenroll	2001	2002	2003	2001	2002	2003
Continue	Undergraduate	From Previous Academic year FA or SP semester: From $P = \text{FA} \cup \text{SP}$	2,813	2,992	3,343	89%	89%	90%
		From P′: Other	336	359	375	11%	11%	10%
Total			*3,149*	*3,351*	*3,718*	*100%*	*100%*	*100%*

To illustrate how our computer algorithm will work, let's consider the mutually exclusive sets created from $P = \text{FA} \cup \text{SP}$. According to Equation 3.1, the number of mutually exclusive subsets we can create is $\text{Max}(P) = 2^2\text{-}1 = 3$, which are defined in Table 4 below. From Table 4, the set A_1 is a subset of students that attend only in the Fall Semester, A_2 is the subset of students that attending in both semesters FA and SP semesters, and A_3 is the set of students that attended only in the Spring Semester. Given A_1, A_2, and A_3 we can combine these sets in different ways in order to obtain a partition of P, see Table 5. Then for each subset that is part of the partition of P, we will use our computer algorithm to explore the historical reenrollment patterns of that set.

Table 4

Max(P) = 3 Mutually Exclusive Subsets of $P = FA \cup SP$

Subset	Previously labeled
$A1 = FA \cap (FA \cap SP)'$	*Fall_Subset*
$A2 = FA \cap SP$	*Intersection*
$A3 = SP \cap (FA \cap SP)'$	*Spring_Subset*

Table 5

Partitioning Methods for $P = FA \cup SP$

Method	Partition of P
1	$A1 \cup A2 \cup A3$
2	$(A1 \cup A2) \cup A3$
3	$A1 \cup (A2 \cup A3)$
4	$(A1 \cup A2 \cup A3)$

In Table 5, each of the partitions contains the same mutually exclusive sets. However, the associations of those mutually exclusive sets differ. For example, Method 1 contains three sets that define the partition of P: A_1, A_2, and A_3. Method 2 and Method

3 contain two sets that create the partition. For partition Method 2 the two sets that define the partition of P are $A_1^* = (A_1 \cup A_2)$ and $A_2^* = A_3$. For partition Method 3, the sets that define the partition of P are $A_1^* = A_1$ and $A_2^* = (A_2 \cup A_3)$. Whereas, Method 4 contains only one set: $A_1^* = (A_1 \cup A_2 \cup A_3)$. Now, for each method and each subset of the partition, our program will explore the enrollment patterns of that subset. The combination of the enrollment patterns of the subsets that defines the partition method will be examined in order to determine which partition method shows the most reliable historical patterns on y_k.

Table 6

Variation in Partitioning

				Return Next Fall (y=1)	
				Year 1	Year 2
Method	Partition	Year 1	Year 2	(rate)	(rate)
Method 1	Fall_Main	91	90	7	11
	$(A1 \cup A2)$			-7.70%	-12.20%
	Spring_Subset	9	10	2	5
	$A3$			-22.20%	-55.50%

The program must choose between all the ways of partitioning P presented in Table 5. The example presented in Table 6 represents one way of partitioning P, but the logical discussion that follows concerning identifying this partition's historical consistency of enrollment can be applied to any of the partitioning methods. Historical consistency relates to the variation that will be introduced by each partitioning method. The partitioning method that introduces the least amount of variation will be deemed the "best" method of partitioning. To get an idea of the partitioning variability that occurs, see Table 6. In Table 6 we can see that out of 91 students in the set Fall_Main = $(A_1 \cup A_2)$, 7 or 7.7% of them returned the following fall semester. If we expect to find consistent historical patterns then we can assume that the following year, in this case Year 2, approximately 7.7% of the students in the subset Fall_Main = $(A_1 \cup A_2)$ for Year 2 will reenroll the following fall. As you can see, this information is known, because we are dealing with historical observations. As it turns out, in Year 2, out of 90 students in the subset Fall_Main = $(A_1 \cup A_2)$, 11 or 12.2% of them returned the following fall semester. Thus, from Year 1 to Year 2 we see variability in the reenrollment rate for the subset Fall_Main = $(A_1 \cup A_2)$. Similarly, from Year 1 to Year 2 we see variability in the reenrollment for the subset Spring_Subset = A_3. In particular, out of 9 students in Spring_Subset = A_3, 2 or (22.2%) reenrolled into the following fall, whereas in Year 2, out of 10 students in Spring_Subset = A_3, 5 or (55.5%) reenrolled into the following fall. Thus, our program needs a way to evaluate historical pattern created by partitioning in order to determine its consistency and compare this partitioning methods consistency to other partitioning methods consistency. To do this we will use a variation of the chi-squared statistic to determine which partitioning method shows the least amount of variation or the most consistent historical patterns.

The typical chi-squared statistic, denoted as χ^2, is calculated by the following formula

$$\chi^2 = \sum \frac{(\text{Observed count } (O) - \text{Expected count } (E))^2}{\text{Expected count } (E)} \tag{4.1}$$

where observed, denoted as O, represents an observed cell count and "expected", denoted as E, represents the expected count for the same cell (2009, Moore, McCabe, Craig). According to Moore, McCabe, and Craig, the *expected* (E) cell count is calculated by

$$E = \frac{\text{Row Total } \times \text{ Column Total}}{\text{Table Total}}. \tag{4.2}$$

To illustrate this we will present an alternative form of the Table 6 information, for Year 1 data only, see Table 7. Notice that under the observed column in Table 3, the information for Year 1 data is exactly the same as Year 1 data in Table 6, but presented in a different format. For each cell in the observed potion of this table there corresponds an expected cell value for the corresponding row and column heading. Thus, for example, 84 students are in the cell corresponding to the students in the subset Fall_Main and no-return heading under observe. Using Equation 4.2, the corresponding expected cell count for Fall_Main and no-return is

$$E = \frac{91 \times 91}{100} = 82.81.$$

The rest of the expected cell counts are calculated in the exact same way. However, in our research we will propose using the chi-statistic, Equation 4.1, but with an alternative to the expected count represented by Equation 4.2, in order to determine the historical variation of a partitioning method.

Table 7

Observed and Expected Number of Students for Year 1

Academic Year 1						
	Observed				Expected	
Partition	no-return	yes-return	Total	Partition	no-return	yes-return
Fall_Main $(A1 \cup A2)$	84	7	91	Fall_Main $(A1 \cup A2)$	82.81	8.19
Spring_subset $A3$	7	2	9	Spring_subset $A3$	8.19	0.81
Total	91	9	100			

To illustrate how we will specify an alternative expected value than Equation 4.2 we will refer to the data in Table 8. Notice that Table 8 considers two consecutive academic years of historical patterns. To discern the two years of academic information we will use appropriate notation. Let $P_1 = FA_1 \cup SP_1$ represent the set of students from academic Year 1 of size N_1, in Table 8 $N_1 = 100$. For any partition method there exists $k \leq Max(P)$ mutually exclusive associations A_1^*, A_2^*, …, A_k^*. In Table 9, the composite set is $P = FA \cup SP$, therefore $Max(P) = 3$, and these mutually exclusive sets are listed in Table 4. In addition, the partition of consideration is Table 9 is $P = (A_1 \cup A_2) \cup A_3$, Method 2 from Table 5, where $k = 2$ such that $A_1^* = (A_1 \cup A_2)$ and $A_2^* = A_3$. Now, for any mutually exclusive association A_j^* of size $N_{1j} \leq N_1$ from academic Year 1 there exists R_1 students who reenroll and R_1' that do not reenroll into the following fall semester. For example, the association $A_1^* = (A_1 \cup A_2)$ in Table 8 has $N_{1j} = 100$, $R_1 = 7$ and $R_1' = 84$. An important calculation we need from academic Year 1 patterns is the enrollment proportion for returning and non-returning students from association A_j^*. In academic Year 1, the proportion of students from A_j^* that reenroll is

$$p_{R_1} = \frac{R_1}{N_{1j}} \tag{4.3}$$

and the proportion of students from A_j^* that do not reenroll is

$$p_{R_1'} = \frac{R_1'}{N_{1j}}. \tag{4.4}$$

In Table 8, academic Year 1 association $A_1^* = (A_1 \cup A_2)$ has $p_{R_1} = \frac{7}{91} = 0.077$ (or 7.7% reenrollment) and $p_{R_1'} = \frac{84}{91} = .923$ (or 92.3% non-reenrollment). For academic Year 2, let $P_2 = FA_2 \cup SP_2$ represent the set of students of size N_2, in Table 8 $N_2 = 100$. Also, the corresponding mutually exclusive association to academic Year 1, the association A_j^* of size $N_{2j} \leq N_2$ from academic Year 2 contains R_2 students who reenroll and R_2' that do not reenroll into the following fall semester. For example, the association $A_1^* = (A_1 \cup A_2)$ in Table 8 has $R_2 = 11$ and $R_2' = 79$. If we expect to find consistent reenrollment patterns from academic Year 1 to academic Year 2, then we expect association A_j^* for academic Year 2 to have a reenrollment proportion similar to p_{R_1} from academic Year 1 and to also have a non-reenrollment proportion similar to $p_{R_1'}$ from academic Year 1. Thus, for any consecutive pair of academic years with association A_j^*, we will use Equation 4.3 to define our expected value of returning students for the most current historical academic year, in this case academic Year 2, as

$$E_{R_2} = N_{2j}(p_{R_1}), \tag{4.5}$$

and using Equation 4.4 our expected value of non-returning students is

$$E_{R_2'} = N_{2j}(p_{R_1'}). \tag{4.6}$$

From Table 8, association $A_1^* = (A_1 \cup A_2)$ for academic Year 2 has an expected reenrollment of $E_{R_2} = N_{21}(p_{R_1}) = 90(0.077) = 6.93$ and an expected non-reenrollment of $E_{R_2'} = N_{21}(p_{R_1'}) = 90(.923) = 83.07$.

Table 8

Historical Enrollment Patterns Under Partition $P = (A_1 \cup A_2) \cup A_3$

		Academic Year 1				Academic Year 2	
	Total	no-return	yes-return		Total	no-return	yes-return
Partition	(N)	(Row %)	(Row %)	Partition	(N)	(Expected)	(Expected)
Fall_Main		84	7	Fall_Main		79	11
$A_1^* = (A1 \cup A2)$	91	-92.30%	-7.70%	$A_1^* = (A1 \cup A2)$	90	-83.07	-6.93
Spring_Subset		7	2	Spring_Subset		5	5
A_2^*	9	-77.80%	-22.20%	A_2^*	10	-7.78	-7.78
Total	100	91	9	Total	100	84	16

Our main goal is to find a way to evaluate historical patterns created by partitioning. For any partition method there exists $k \le Max(P)$ mutually exclusive associations A_1^*, A_2^*, …, A_k^* that define the partition method: P = $A_1^* \cup A_2^* \cup \cdots \cup A_k^*$. To evaluate the historical consistency of the patterns found in A_j^*, j $\in$ {1, 2, ..., k}, for any consecutive pair of academic years, we will use Equation 4.5 and 4.6 to obtain the following chi-squared statistic

$$\chi^2_{A_j^*} = \frac{\left(R_2 - E_{R_2}\right)^2}{E_{R_2}} + \frac{\left(R_2' - E_{R_2'}\right)^2}{E_{R_2'}}. \tag{4.7}$$

Notice that the numerator of Equation 4.6 takes the actual historical reenrollment in Year 2 and squares the distance that value is from the expected reenrollment you would get using the previous year's enrollment proportion. Thus, the more consistency we have in actual enrollment proportions from Year 1 to Year 2, the smaller the chi-squared value in Equation 4.6. For example, if Year 2 actual reenrollment proportion, $p_{R_2} = \frac{R_2}{N_{2j}}$, is equivalent to Equation 4.3 and if Year 2 actual non-reenrollment proportion, $p_{R_2'} = \frac{R_2'}{N_{2j}}$, is equivalent to Equation 4.4, then Equation 4.6 will be zero, which indicates perfect reenrollment consistency for association A_j^*, or no variation in reenrollment patterns from academic Year 1 to academic Year 2. To evaluate a partition method P = $A_1^* \cup A_2^* \cup \cdots \cup A_k^*$ for historical consistency on reenrollment patterns of y_k, we will use Equation 4.6 to calculate the following chi-squared value

$$\chi_P^2 = \chi_{A_1^*}^2 + \chi_{A_2^*}^2 + \dots + \chi_{A_k^*}^2 . \tag{4.8}$$

For example, using Equation 4.6 the association $A_1^* = (A_1 \cup A_2)$ from Table 8 will generate chi-squared value

$$\chi_{A_1^*}^2 = \frac{(11-6.93)^2}{6.93} + \frac{(79-83.07)^2}{83.07} = 2.59.$$

In addition, the association $A_2^* = A_3$ from Table 8 will generate a chi-squared value

$$\chi_{A_2^*}^2 = \frac{(5-7.78)^2}{7.78} + \frac{(5-7.78)^2}{7.78} = 1.99.$$

Thus, using Equation 4.7, the historical consistency of the partition method presented in Table 8 is

$$\chi_P^2 = 2.59 + 1.99 = 4.58.$$

Now, suppose we have competing partition methods $P_1, P_2, \dots, P_n$, then for each partition method, using Equation 4.7, we can calculate corresponding chi-squared values $\chi_{P_1}^2$, $\chi_{P_2}^2$, …, $\chi_{P_n}^2$. Our *decision rule* on which partition method is best will be determined by

$$\text{Best Partition} = \text{Min}\left(\chi_{P_1}^2, \chi_{P_2}^2, \dots, \chi_{P_n}^2\right). \tag{4.9}$$

In our study we examine three years of historical enrollment patterns for composite snap shot $P = FA \cup SP$. The Academic years of data we will explore for historical consistency on y_k, as determined by Equation 4.7, are academic Year 1 = {Fall 2001, Spring 2002}, academic Year 2 = {Fall 2002, Spring 2003}, and academic Year 3 = {Fall 2003, Spring 2004}. Having three years of historical patterns to examine is a simple extension of the discussion we had for two academic years of historical patterns presented in Table 8. In general, for any consecutive pair of academic years with association A_j^*, for some partition method, we will calculate a chi-squared statistic presented by Equation 4.6. To determine the overall value of or Equation 4.6 will require adding the corresponding values of Equation 4.6 for each consecutive pair of academic years. For example, suppose we are considering the association A_j^*, for some partition method, over three years of historical patterns. First, we will explore the historical patterns from academic Year 1 to academic Year 2 for association A_j^* and determine the appropriate chi-squared value represented by Equation 4.6, which we will denote as $\chi^2_{(12)A_j^*}$. Similarly, we explore the historical patterns from academic Year 2 to academic Year 3 for the same association A_j^* and determine the appropriate chi-squared value represented by Equation 4.6, which we will denote as $\chi^2_{(23)A_j^*}$. The overall chi-squared value, Equation 4.6, over three years of historical patterns for association A_j^* is $\chi^2_{A_j^*}$ = $\chi^2_{(12)A_j^*} + \chi^2_{(23)A_j^*}$. Once $\chi^2_{A_j^*}$ is obtained then all the subsequent calculations needed in order to determine the best partition follow. Notice that this logical procedure can easily be extended for historical patterns extended beyond three years of historical patterns. Table 9 shows the results of Equation 4.7 from using three years of historical reenrollment patterns on composite data $P = FA \cup SP$ for all the partition methods listed in Table 5. From Table 9 you can see that hat the decision rule would select partition Method 4 as the best then the second best partition method is Method 2. We now extend our discussion to include independent variables defined in Equation 3.4 that will exist for the individuals in $P = FA \cup SP$ and combine them into our program to obtain stratification within the association A_j^* to possibly obtain even more consistent enrollment patterns.

Table 9

Chi-square Values for Different Partitioning Methods

Partition Method	Chi Square	Best
1 - $A1 \cup A2 \cup A3$	284.0888	4
2 - $(A1 \cup A2) \cup A3$	268.958	2
3 - $A1 \cup (A2 \cup A3)$	282.2263	3
4 - $(A1 \cup A2 \cup A3)$	266.8814	1

For institutions of higher education there is typically useful information $\mathbf{x}_k$ from Equation 3.4 that is known concerning their students in the composite data set $P = FA \cup SP$. In our program we will also examine the historical enrollment patterns of the partition methods by stratifying them according to categorical variables within $\mathbf{x}_k$. In this study some of the variables we consider using for stratification are $\mathbf{x}_k$ = [age$_k$, ethnicity$_k$, GPA$_k$]. The idea of stratification using categorical variables is important for our program since we will be able to produce frequency tables like Table 8. To integrate these categorical variables within the context of calculating the appropriate chi-squared values in order to determine is very straight forward. Suppose we are examining some partition $\mathrm{P} = A_1^* \cup A_2^* \cup \cdots \cup A_k^*$ and wish to stratify that partition according to a categorical variable with p categories. Then for any for any consecutive pair of academic years with association A_j^*, we will calculate a chi-squared statistic presented by Equation 4.6 for each stratum. For example, for strata 1 under association A_j^*, we will calculate the appropriate chi-squared value represented by Equation 4.6, which we will denote as $\chi^2_{A^*_{j(1)}}$. Our program will do this for each of the p-strata and will generate the overall chi-squared value, Equation 4.6, for association A_j^* by the following summation: $\chi^2_{A_j^*} = \chi^2_{A^*_{j(1)}} + \chi^2_{A^*_{j(2)}} + \cdots + \chi^2_{A^*_{j(p)}}$. An interesting dilemma arises when you consider injecting strata into the associations that define the partition. If an association A_j^* has information $\mathbf{x}_k$ from different time periods T = t, then the program will tag that set with the information from the different time periods and calculate separate chi-squared values, Equation 4.6, using the categorical variable for the each tagged time t . Table 11, presents the chi-squared results of Equation 4.7 for the partitions using categorical variables for stratification. The union of sets that were made at different times then the program needs to consider tagging the set. For example, partition Method 2 contains the mutually exclusive set $A_2 = FA \cap SP$ which contains $\mathbf{x}_k$ information from the fall semester as well as $\mathbf{x}_k$ information from the spring semester. However, since we associated that set with $A_1 = FA \cap (FA \cap SP)'$ in Method 2, then we tagged A_2 with the information $\mathbf{x}_k$ contained during the time of the fall semester. Similarly, Method 4 contains $A_2 = FA \cap SP$ so Table 11 shows two results for that partitioning method, one when A_2 is tagged with spring information and the other when A_2 is tagged with fall information. One of the most interesting things to note from comparing the chi-squared values from Table 9 to Table 11 is that in every case, adding a variable to stratify the partition facilitates finding more consistent patterns of y_k. In other words, the chi-squared values, calculated using Equation 4.7, in Table 11 are smaller than those in Table 9.

Table 10

Some Categorical Variables Used

VARIABLE	DEFINITION	CODING
Class	Class level	C1 (Doctorate Degree) C2 (Freshman) C3 (Junior) C4 (Master) C5 (Post Baccalaureate) C6 (Sophomore) C7 (Senior) C8 (Sophomore)
Ethnicity	Ethnicity	E1 (white, non-hispanic) E2 (black, non-hispanic) E3 (hispanic) E4 (asian, amer./pac. isl.) E5 (Amer. Indian/alaskan) E6 (international) E7 (other)
Level	Class Level	L1 (Undecided) L2 (Undergraduate) L3 (Post Baccalaureate) L4 (Graduate Nursing) L5 (Graduate Master) L6 (Graduate Doctorate) L7 (Graduate Certificate)

Table 11

Partitions Using Stratifications by Categorical Variables

Partition Method	Stratification Variable	Chi Square Using Stratification	Best
1 - $A_1 \cup A_2 \cup A_3$ A_2 Tagged by Spring	Year*Level	250.1027	6
1 - $A_1 \cup A_2 \cup A_3$ A_2 Tagged by Fall	Year*Level	243.9577	4
2 - $(A_1 \cup A_2) \cup A_3$	Year*Level	200.1027	3
3 - $A_1 \cup (A_2 \cup A_3)$	Year*Level	248.2425	5
4 - $(A_1 \cup A_2 \cup A_3)$- A_2 Tagged by Spring	Year*Level	198.3771	2
4 - $(A_1 \cup A_2 \cup A_3)$- A_2 Tagged by Fall	Year*Level	191.1046	1

Although each partition method of P = FA $\cup$ SP is composed of the same individuals, Table 11 shows that enrollment patterns differ when you partition the set and stratify the set in different ways. In Chapter V will illustrate that using the best partition of P = FA $\cup$ SP, as defined by our decision rule 4.8, makes a difference in terms of the accuracy of predictions. Our contention is that modeling the best partitioning method, as defined by 4.8, for prediction will yield better predictive results than modeling other partitioning methods for prediction, even though all partition methods have the same individuals.

CHAPTER V

DEVELOPMENT OF PREDICIVE MODELS

Previously we mentioned that, good data in and of itself is not inherently informative. In this regard, Chapter IV discussed how our program can iteratively explore or mine composite data sets in order to obtain useful information on historical snap shots of data. Particularly, we have created a program that uses our derived decision rule, Equation 4.8, as a pre-modeling technique in order to determine information on potential usable and reliable patterns that will facilitate predictive modeling or even simple reporting. For predictive modeling to be effective, the coefficients that define the model need to be sustained in future data related to the variable of interest. Thus, since our program identifies the most consistent historical patterns of y_k by partitioning and stratifying the snap shots of data, then modeling the variable of interest using the best partitions, assuming these historical patterns persist, should produce the most accurate predictions on observed y_k In this chapter we will construct an appropriate predictive model on the top three partitions determined in Chapter IV and compare how well they predict y_k on future data that was not used to make the model.

For the dichotomous variable y_k, a researcher could specify various models to aid them in their predictive endeavors. Our contention is that for any specified modeling technique on the random variable y_k where $k \in P$, predictive accuracy will benefit by using the best partition of P. To illustrate this, we will construct a model on random variable y_k using one historical academic year of data, $P_{2002} = FA \cup SP$. Now, historical academic year data $P_{2002} = FA \cup SP$ will have observed values of y_k, where $y_k = 1$ if student k reenrolled in Fall 2003 and $y_k = 0$ if student k did not reenroll in Fall 2003. In this regard, we can associate for each individual $k \in P_{2002}$ a set of p independent variables $\mathbf{x}_k = \left\{x_{1k}, x_{2k}, \ldots x_{pk}\right\}$, to explain y_k. For each association A_i^* from the partition $P_{2002} = A_1^* \cup A_2^* \cup \cdots \cup A_k^*$ and for $k \in A_i^*$, we will use SAS software to specify a logistic regression model of the form

$$L_k = \beta_0 + \beta_1 x_{1k} + \beta_2 x_{2k} + \ldots \beta_p x_{pk}, \tag{5.1}$$

where β_0, $\beta_1, \ldots,$ β_p represent the coefficient values, which are determined through maximum likelihood estimation by SAS. For a more detailed discussion on logistic regression see Ahluwalia (2006) and Ingram (2008).

Given a specified "best" partition for academic year P_{2002}, we want to explore the equivalent partitioning method, $P_{2003} = A_1^* \cup A_2^* \cup \cdots \cup A_k^*$, in the subsequent academic year 2003. In particular, for partitioning method, $P_{2003} = A_1^* \cup A_2^* \cup \cdots \cup A_k^*$, we want to predict the total number of students from each association A_i^*

$$T_{A_i^*} = \sum_{k \in A_i^*} y_k \,. \tag{5.2}$$

From Equation 5.2, we will extend the discussion to the total reenrollment for partition $P_{2003} = A_1^* \cup A_2^* \cup \cdots \cup A_k^*$, which is

$$T_{P_{2003}} = T_{A_1^*} + T_{A_2^*} + \cdots + T_{A_k^*} \,. \tag{5.3}$$

Given academic year 2003 partition $P_{2003} = A_1^* \cup A_2^* \cup \cdots \cup A_k^*$, for each $k \in A_i^*$ using corresponding independent variables $\mathbf{x}_k$ into Equation 5.1, we can calculate the probability that an individual k will reenroll by using the following equation

$$P(y_k = 1) = \frac{\exp(L_k)}{1 + \exp(L_k)} \tag{5.4}$$

Thus, our prediction for Equation 5.2 is

$$\hat{T}_{A_i^*} = \sum_{k \in A_i^*} P(y_k = 1) \,. \tag{5.5}$$

Accordingly, using Equation 5.5, the total predicted number of re-enrolling students from P_{2003}, under the appropriate partition, into the following Fall semester (Fall 2004), is

$$\hat{T}_{Partition} = \sum_{i=1}^{k} \hat{T}_{A_i^*} . \tag{5.6}$$

Prediction for the best, Method-1, partition: $P = (A_1 \cup A_2 \cup A_3)$-Tagged by Fall

For the best partition, $P = (A_1 \cup A_2 \cup A_3)$*-Tagged by Fall*, identified in Chapter IV, Equation 5.1 for association $A_1^* = (A_1 \cup A_2 \cup A_3)$*-Tagged by Fall* coming from P_{2002} *is*

$$\begin{aligned} L_k = {} & -.3562+(-.351*E1(\text{white, non-hispanic}))+ \\ & +(-.285*E2(\text{black, non-hispanic}))+(-.1742*E3(\text{hispanic}))+ \\ & +(-.3186*E4(\text{asian, amer./pac. isl.}))+(-.2639*E5(\text{amer. indian/alaskan}))+ \\ & +(-.106*E6(\text{international}))+(-.4094*C1(\text{Docorate Degree}))+ \\ & +(-.3024*C2(\text{Freshman}))+(.3457*C3(\text{Junior}))+(-.7094*C4(\text{Master}))+ \\ & +(-.7773*C5(\text{Post Baccalaureate}))+(-1.1282*C7(\text{Senior}))+ \\ & +(-.1685*L1(\text{Undecided}))+(1.2675*L2(\text{Undergraduate}))+ \\ & +(1.4166*L3(\text{Post Baccalaureate}))+(-.00973*L4(\text{Graduate Nursing}))+ \\ & +(1.4663*L5(\text{Graduate Master}))+(1.6821*L6(\text{Graduate Doctorate})) \end{aligned} \tag{5.7}$$

The SAS coefficient output for Equation 5.7 above is found in Table 21 Appendix C. In addition, practical interpretations of the beta coefficients in Equation 5.7 can be obtained from the odds ratios table (see Table 22 from Appendix C).

In order to assess the fit of our model we used a Hosmer and Lemeshow goodness-of-fit test, which tests the hypothesis that there are no differences between observed response variable values, y_k, from the training data set A_i^* coming from P_{2002}, and the estimated response values calculated by using model 5.7. Table 12 shows that we failed to reject the above hypothesis which indicates the model fits the data well.

Table 12

Hosmer and Lemeshow Goodness-of-Fit Test

Hosmer and Lemeshow Goodness-of-Fit Test		
Chi-Square	DF	Pr > ChiSq
1.5419	7	0.9808

Now using Equation 5.4 and 5.5, the predicted reenrollment from association $A_1^* = (A_1 \cup A_2 \cup A_3)$*-Tagged by Fall* coming from $P_{2003} = (A_1 \cup A_2 \cup A_3)$ is found in Table 13. In Using Equation 5.6, Table 13 shows the total predicted re-enrollment using this partition method $P_{2003} = (A_1 \cup A_2 \cup A_3)$, identified as the best partition Method 1, is $\hat{T}_{Method-1} =$

6,853. The actual reenrollment total, Equation 5.3, coming from $P_{2003} = (A_1 \cup A_2 \cup A_3)$ is $T_{P_{2003}} = 6{,}577$. Thus, using partition Method 1 the prediction was off by 276.

Table 13

Predicted Re-enrollment from $A_1^* = (A_1 \cup A_2 \cup A_3)$*-Tagged by Fall*

	Fall 2004 REENROLLEMENT PREDICTION		
LEVEL	A_1^*	Fall 2004 Predicted Re-enroll	Fall 2004 Actual Re-enroll
10% TO < 20%	882	171	201
20% TO < 30%	466	104	62
30% TO < 40%	2,216	811	786
40% TO < 50%	798	360	331
50% TO < 60%	5,372	2,856	2,640
60% TO < 70%	2,242	1,442	1,448
70% TO < 80%	1,539	1,110	1,109
Total	*13,515*	*6,853*	*6,577*

Next we will explore the predictive accuaracy of the second best partion method that was identified in Chapter IV.

Prediction for second best, Method 2, partition: P = (($A_1 \cup A_2$*)*$\cup A_3$*)-Tagged by Fall*

For the best partition, $P = (A_1 \cup A_2) \cup A_3$-*Tagged by Fall*, identified in Chapter IV, Equation 4.1 for association $A_1^* = (A_1 \cup A_2)$ and $A_2^* = A_3$ coming from P_{2002} *are*

$$\begin{aligned} L_{FM} = {} & -.9261+(.0867*E1(\text{white, non-hispanic}))+ \\ & +(.0965*E2(\text{black, non-hispanic}))+(.2407*E3(\text{hispanic}))+ \\ & +(.0878*E4(\text{asian, amer./pac. isl.}))+(.0652*E5(\text{amer. indian/alaskan}))+ \\ & +(.2907*E6(\text{international}))+(.2153*C1(\text{Docorate Degree}))+ \\ & +(-.4205*C2(\text{Freshman}))+(.2293*C3(\text{Junior}))+(-.0364*C4(\text{Master}))+ \\ & +(.3044*C5(\text{Post Baccalaureate}))+(-1.1952*C7(\text{Senior}))+ \\ & +(.0180*L1(\text{Undecided}))+(1.9091*L2(\text{Undergraduate}))+ \\ & +(.8759*L3(\text{Post Baccalaureate}))+(.0323*L4(\text{Graduate Nursing}))+ \\ & +(1.2711*L5(\text{Graduate Master}))+(1.6227*L6(\text{Graduate Doctorate})) \end{aligned} \tag{5.8}$$

Similarly, for Spring subset we will have:

$$\begin{aligned} L_{SS} = {} & -1.2973+(.3988*E1(\text{white, non-hispanic}))+ \\ & +(.5675*E2(\text{black, non-hispanic}))+(.3811*E3(\text{hispanic}))+ \\ & +(.8082*E4(\text{asian, amer./pac. isl.}))+(-11.5675*E5(\text{amer. indian/alaskan}))+ \\ & +(-.7167*C1(\text{Docorate Degree}))+(-.1567*C2(\text{Freshman}))+ \\ & +(.5327*C3(\text{Junior}))+(-2.0882*C4(\text{Master}))+ \\ & +(-.9975*C5(\text{Post Baccalaureate}))+(-.2052*C7(\text{Senior}))+ \\ & +(-1.3572*L2(\text{Undergraduate}))+(.1752*L4(\text{Graduate Nursing}))+ \\ & +(.7330*L5(\text{Graduate Master})) \end{aligned} \tag{5.9}$$

The SAS coefficient output for Equation 5.8 and 5.9 above is found in Table 24 and Table 27 from Appendix C. In addition, practical interpretations of the beta coefficients in Equation 5.7 can be obtained from the odds ratios table (see Table 25 and Table 28 from Appendix C).

In order to assess the fit of our model we used a Hosmer and Lemeshow goodness-of-fit test, which tests the hypothesis that there are no differences between observed response variable values, y_k, from the training data set A_i^* coming from P_{2002}, and the estimated response values calculated by using model 5.8 and 5.9. Table 14 and Table 15 shows that we failed to reject the above hypothesis which indicates the model fits the data well.

Table 14

Hosmer and Lemeshow Goodness-of-Fit Test for A_1^*

Hosmer and Lemeshow Goodness-of-Fit Test		
Chi-Square	DF	Pr > ChiSq
0.9028	7	0.9962

Table 15

Hosmer and Lemeshow Goodness-of-Fit Test for A_2^*

Hosmer and Lemeshow Goodness-of-Fit Test		
Chi-Square	DF	Pr > ChiSq
5.1905	7	0.6367

Now using Equation 5.4 and 5.5, the predicted reenrollment from association $A_1^* = (A_1 \cup A_2)$ and $A_2^* = A_3$, A_2-*Tagged by Fall* coming from $P_{2003} = (A_1 \cup A_2) \cup A_3$ is found in Table 14. In using Equation 5.6, Table 16 shows the total predicted re-enrollment using this partition method $P_{2003} = (A_1 \cup A_2\ ) \cup A_3$, identified as the best partition, Method 2, is $\hat{T}_{Method-2} = 6{,}892$. The actual reenrollment total, Equation 5.3, coming from $P_{2003} = (A_1 \cup A_2) \cup A_3$ is $T_{P_{2003}} = 6{,}577$. Thus, using partition Method 2 the prediction was off by 315.

Table 16

Predicted Reenrollment from $A_1^* = (A_1 \cup A_2) \cup A_3 = A_1^* \cup A_2^*$ *, where* A_2 *is Tagged by Fall*

2004 CONTINUING REENROLLEMENT PREDICTION							
LEVEL	A_1^*	Fall Main 2004 Predicted Re-enroll	Fall Main 2004 Actual Re-enroll	LEVEL	A_2^*	Spring Subset 2004 Predicted Re-enroll	Spring Subset 2004 Actual Re-enroll
10% TO < 20%	.	.	.	10% TO < 20%	2,108	168	185
20% TO < 30%	3	0	1	20% TO < 30%	626	87	77
30% TO < 40%	146	41	47	30% TO < 40%	28	7	3
40% TO < 50%	730	329	186	40% TO < 50%	2	1	0
50% TO < 60%	1,668	783	742	50% TO < 60%	1	0	0
60% TO < 70%	3,561	2,100	1,983	60% TO < 70%	.	.	.
70% TO < 80%	1,565	1,025	1,006	70% TO < 80%	.	.	.
80% TO < 90%	2,866	2,181	2,184	80% TO < 90%	.	.	.
90% TO < 100%	211	171	163	90% TO < 100%	.	.	.
Total	*10,750*	*6,629*	*6,312*	*Total*	*2,765*	*263*	*265*

Next we will explore the predictive accuaracy of the second best partion method that was identified in Chapter IV.

Prediction for third best partition, Method 3: $P = A_1 \cup (A_2 \cup A_3)$ - Tagged by Spring

For the best partition, $P = A_1 \cup (A_2 \cup A_3)$ *-Tagged by Spring*, identified in Chapter IV, Equation 4.1 for association $A_1^* = A_1$ and $A_2^* = (A_2 \cup A_3)$ coming from P_{2002} *are*

L_{FS}= 9.5025+(-.0161*E1(white, non-hispanic))+
+(-.0108*E2(black, non-hispanic))+(.1414*E3(hispanic))+
+(-.023*E4(asian, amer./pac. isl.))+(-.1474*E5(amer. indian/alaskan))+
+(.0865*E6(international))+(-10.3026*C1(Docorate Degree))+
+(-.2135*C2(Freshman))+(.1271*C3(Junior))+(-10.4818*C4(Master))+ (5.10)
+(.5814*C5(Post Baccalaureate))+(-.5463*C7(Senior))+
+(-10.5183*L1(Undecided))+(-8.6189*L2(Undergraduate))+
+(-9.3578*L3(Post Baccalaureate))+(-.00503*L4(Graduate Nursing))+
+(1.9722*L5(Graduate Master))+(2.0018*L6(Graduate Doctorate))

Similarly, for Spring Main subset we will have:

L_{SS}= =-.5396+(-.2735*E1(white, non-hispanic))
+(-.2078*E2(black, non-hispanic))+(-.0354*E3(hispanic))
+(-.1125*E4(asian, amer./pac. isl))
+(.1280*E5(amer. Indian/alaskan))
+(10.8663*C1(Doctorate Degree))+(-.5278*C2(Freshman))
+(-.343*C3(Junior))+(-.8756*C4(Master)) (5.11)
+(-.604*C5(Post Baccalaureate))+(-1.7012*C7(Senior))
+(-10.6609*L1(Undecided))+(1.4506*L2(Undergraduate))
+9(.9302*L3(Post Baccalaureate))+(.3201*L4(Graduate Nursing))
+(1.19*L5(Graduate Master))+(-9.6709*L6(Graduate Doctorate))

The SAS coefficient output for Equation 5.10 and 5.11 above is found in Table 21 from Appendix C. In addition, practical interpretations of the beta coefficients in Equation 5.10 and 5.11 can be obtained from the odds ratios table (see Table 21 and Table 22 from Appendix C).

In order to assess the fit of our model we used a Hosmer and Lemeshow goodness-of-fit test, which tests the hypothesis that there are no differences between observed response variable values, y_k, from the training data set A_i^* coming from P_{2002}, and the estimated response values calculated by using model 5.10 and 5.11. Table 17 and Table 18 shows that we failed to reject the above hypothesis which indicates the model fits the data well.

Table 17

Hosmer and Lemeshow Goodness-of-Fit Test for A_1^*

Hosmer and Lemeshow Goodness-of-Fit Test		
Chi-Square	DF	Pr > ChiSq
2.9466	7	0.8899

Table 18

Hosmer and Lemeshow Goodness-of-Fit Test for A_2^*

Hosmer and Lemeshow Goodness-of-Fit Test		
Chi-Square	DF	Pr > ChiSq
5.1905	7	0.6367

Now using Equation 5.10 and 5.11, the predicted reenrollment from association $A_1^* =$ $A_1 \cup (A_2 \cup A_3)$*-Tagged by Spring* coming from $P_{2003} = A_1 \cup (A_2 \cup A_3)$ is found in Table 17 and Table 18. In Using Equation 5.10 and 5.11, Table 19 shows the total predicted re-enrollment using this partition method $P_{2003} = A_1 \cup (A_2 \cup A_3)$, identified as the best partition, Method 3, is $\hat{T}_{Method-3} = 6{,}937$. The actual reenrollment total, Equation 5.10 and 5.11, coming from $P_{2003} = (A_1 \cup A_2 \cup A_3)$ is $T_{P_{2003}} = 6{,}577$. Thus, using partition Method 3 the prediction was off by 360.

Table 19

Predicted Reenrollment from $A_1^* = A_1 \cup (A_2 \cup A_3) = A_1^* \cup A_2^*$ *, where* A_2 *is Tagged by Spring*

2004 CONTINUING REENROLLEMENT PREDICTION							
LEVEL	A_1^*	Fall Subset 2004 Predicted Re-enroll	Fall Subset 2004 Actual Re-enroll	LEVEL	A_2^*	Spring Main 2004 Predicted Re-enroll	Spring Main 2004 Actual Re-enroll
10% TO < 20%	3	1	1	10% TO < 20%	21	0	5
20% TO < 30%	563	153	139	20% TO < 30%	145	24	29
30% TO < 40%	87	26	16	30% TO < 40%	2,144	529	487
40% TO < 50%	1	0	1	40% TO < 50%	2,909	1,389	1,027
50% TO < 60%	189	110	111	50% TO < 60%	718	311	274
60% TO < 70%	1,003	666	677	60% TO < 70%	2,187	1,245	1,245
70% TO < 80%	2,173	1,585	1,568	70% TO < 80%	1,217	789	756
80% TO < 90%	3	2	1	80% TO < 90%	149	106	105
90% TO < 100%	-	-	-	90% TO < 100%	2	2	2
Total	*4,022*	*2,543*	*2,514*	*Total*	*9,493*	*4,394*	*3,931*

Analizing the best three partitioning methods, as identified in Chapter IV, we have observed that using the same prediction model, in our case the logistic regression, the closest prediction to the actual reenrollement was obtained with the best partion method, Method 1. More than that, the second best partition, Method 2, is predicting less accurate than Method 1, and the third best partion, Method 3, is less accurate than Method 2. In general we see that the order of the best historical patterns identified by our decison rule, are generating accurate prediction in the same order. Thus, our program provides a pre-modeling tool that can help identify the best patterns to model in order to obtain more accruate pedictions.

CHAPTER VI

CONCLUSION

The purpose of this paper was to illustrate that exploration of historical composite snap shots of data, Definition 1, is an important pre-modeling technique that can facilitate finding consistent patterns relevant to a dichotomous dependent variable of interest. Using reenrollment, Equation 3.1, as the dependent variable of interest and composite data set made from fall and spring snap shots of data, see Figure 1, we have created an independent SAS program that can analyze various historical partitions as defined by Definition 2 of the composite snap shots of data. Our program can determine the best, as defined by Equation 4.8, historical patterns created form partitioning and stratification of the historical data. For each of the top three partitions we modeled the dependent variable using appropriate logistic regression models. Then for each partition using the respective logistic regression models we were able to obtain predicted total reenrollments on future data, which was not used to create the logistic models. The results indicated that predictive modeling on the best partition method determined from exploring historical data yielded superior predictions. These superior predictions occurred because the historical best patterns due to partitioning evidently were sustained in future data. In general, we illustrated that modeling the same data, but with different partitions does make a difference in predicting a dependent variable of interest.

For future research we would like to investigate the sampling distribution of the chi-square value we specified in Equation 4.6. We note that the chi-squared statistic we specified for each partition, Equation 4.7, is similar to the Hosmer and Lemeshow statistic for goodness of fit found in Logistic regression. Our chi-squared statistic is trying to see which historical patterns, if you will, fit together most consistently. The main difference is that the Hosmer and Lemeshow statistic uses a fitted logistic regression model on current data to specify appropriate decide probability intervals that are used to determine an expected value (E) for chi-squared Equation 4.1, whereas our chi-squared expected values are determined by historical patterns.

REFERENCES

Goenner, Cullen F., and Kenton Pauls. 2006. A predictive model of inquiry to enrollment. *Research in Higher Education* 47, no. 8. http://ezproxy.twu.edu:2096/content/m513t0512113126h/fulltext.html (accessed June 29, 2008).

Hedeker, Don. 2003. A mixed-effects multinomial logistic regression model. *Statistics in Medicine* 22: 1433-1446.

Hosmer, David W. and Stanley Lemeshow. 2000. *Applied logistic regression.* New York: John Wiley and Sons.

Lawrence, Jean M., Debra A. Standiford, Beth Loots, Georgeanna J. Klingensmith, Desmond E. Williams, Andrea Ruggiero, Angela D. Liese, Ronny A. Bell, Beth E. Waitzfelder, and Robert E. McKeown. 2006. Prevalence and correlates of depressed mood among youth with diabetes: The search for diabetes in youth study. *Pediatrics* 117: 1348-1358.

Long, J. Scott. 1997. *Regression models for categorical and limited dependent variables.* Thousand Oaks: Sage Publications, Inc.

Morley, Kathleen. 2000. The enrollment prediction model-How did we do? Paper presented at the annual conference of the Southern Association for Institutional Research, Myrtle Beach, SC. October.

Olson, Gary A. 2005. Predicting undergraduate mathematics success: An analysis of UCDHSC placement methods and the Accuplacer exam. Master's thesis, University of Colorado at Denver.

O'Shea, T. Michael, Jamanadas M. Kothadia, Kurt L. Klinepeter, Donald J. Goldstein, Barbara G. Jackson, and R. Grey Weaver. 1999. Randomized placebo-controlled trial of a 42-day tapering course of Dexamethasone to reduce the duration of ventilator dependency in very low birth weight infants: Outcome of study participants at 1-year adjusted age. *Pediatrics* 104: 15-21.

Powers, A. Daniel and Yu Xie. 2000. *Statistical methods for categorical data analysis.* San Diego: Academic Press.

Sedwick, Susan Wyatt, D. Stanley Carpenter, Nestor W. Sherman, and Alan Tipton.2001. Measuring the efficacy of telemarketing in student recruitment. *C & U Journal*, Fall: 23-27.

Stokes, Maura E., Charles S. Davis, and Gary G. Koch. 2001. *Categorical data analysis using the SAS system.* Cary, NC: SAS Institute Inc.

TWU Academic Advising Center. Math Placement Test. Texas Woman's University. http://www.twu.edu/aac/math-placement.asp (accessed June 25, 2008).

TWU Factbook. Texas Woman's University. http://www.twu.edu/irp/web/factbook.htm (accessed June 25, 2008).

Zamirski, Eva Beata. 2006. A statistical study of students' performances on entrance exams as predictors of their success in college algebra. Master's thesis. Texas Woman's University.

Zullig, Keith J., Robert F. Valois, E. Scott Huebner, John E. Oeltmann, and J. Wanzer Drane. 2001. Relationship between perceived life satisfaction and adolescents' substance abuse. *Journal of Adolescent Health* 29, no. 4 (October): 279-288. http://www.sciencedirect.com/science/article/B6T80-4434R66-7/1/46434f9f94b873bd5df6fe69fdbea82f (accessed June 28, 2008).

APPENDIX A

List of Independent Variables

Table 20

List of Independent Variables

VARIABLE	DEFINITION	CODING
AGEVAL	Age categories	(0) age value is missing (1) age <25 (2) 25<=age<45 (3) 45<=age
GPAVAL	Average score of GRE, SAT or GMAT	(gmat score) if exists ((gre_q+gre_v)/2) if gmat score is missing ((sat_v+sat_m)/2) if gmat score is missing and gre score is missing too
GPAENTRY	Average score categories	(0) missing (1) gpaval<450 (2) 450<=gpaval
CUMGPA	Cumulative GPA	(0) missing (1) 2<=gpa<3 (2) 3<=gpa
GPA	GPA categories	(0) missing (1) 2<=cumgpa<3 (2) 3<=cumgpa

APPENDIX B

SAS Code

SAS Code

This is the code used to complete this analysis for computing the chi square.

```
* Telling SAS where to find data*

libname FALL 'c:\Thesis\Fall\DATA\';

libname SPRING 'c:\Thesis\Spring\DATA';

libname ALL_SEM 'c:\Thesis\AllSemesters';

libname year99 'c:\Thesis\AllSemesters\Years\99';

libname year00 'c:\Thesis\AllSemesters\Years\00';

libname year01 'c:\Thesis\AllSemesters\Years\01';

libname year02 'c:\Thesis\AllSemesters\Years\02';

libname year03 'c:\Thesis\AllSemesters\Years\03';

libname year04 'c:\Thesis\AllSemesters\Years\04';


*Merging data sets to form one dataset with all years*

*for spring00*;

data project.tmp;

set project.sp_00;

run;

proc sort data=project.tmp out=project.tmp2;

      by id_bog;

run;
```

```
data project.tmp_id;
set project.tmp2 (keep=id_bog term stu_level class);
by id_bog;
if first.id_bog;
run;

*Creating a unique ID in the new dataset*
proc sort data=project.tmp_id out=project.tmp;
      by id_bog;
run;
data all_sem.springID;
set project.tmp;
run;

*Deleting temporary files*
proc datasets library=project;
  delete tmp;
  delete tmp2;
  delete tmp_id;
  delete tmp_id_spring01;
  delete tmp_id_spring02;
  delete tmp_id_spring03;
  delete tmp_id_spring04;
  delete tmp_id_spring05;
run;

*Creating and appending data sets for fall semester*
data FALL.tmp;
set FALL.fa_99;
run;
proc sort data=FALL.tmp out=FALL.tmp2;
      by id_bog term;
run;
data FALL.tmp_id;
set FALL.tmp2 (keep=id_bog term stu_level class);
by id_bog;
if first.id_bog;
run;
*fall00*;
data FALL.tmp;
set FALL.fa_00;
run;
proc sort data=FALL.tmp out=FALL.tmp2;
      by id_bog;
```

```
run;

data FALL.tmp_id_fall00;

set FALL.tmp2 (keep=id_bog term stu_level class);

by id_bog;

if first.id_bog;

run;


proc append base=FALL.tmp_id data=FALL.tmp_id_fall00;

run;


*Adding returning or not variable*

* fa99*;

data ALL_SEM.listed_once;

set ALL_SEM.uniqueID;

by id_bog;

if first.id_bog;

drop stu_level class term;

run;

data ALL_SEM.tmp;

set ALL_SEM.uniqueID;

by id_bog;

if term='99/FA';
```

```
lvl_99FA=stu_level;

cls_99FA=class;

ret_99FA='NO';

drop stu_level class term;

run;

data ALL_SEM.listed_once;

merge ALL_SEM.listed_once ALL_SEM.tmp;

by id_bog;

run;

*sp00*;

data ALL_SEM.tmp;

set ALL_SEM.uniqueID;

by id_bog;

if term='00/SP';

lvl_00SP=stu_level;

cls_00SP=class;

ret_00SP_tmp='NO'; *default=NO*;

drop stu_level class term;

run;

data ALL_SEM.listed_once;

merge ALL_SEM.listed_once ALL_SEM.tmp;

by id_bog;
```

```
run;

data ALL_SEM.listed_once;

set ALL_SEM.listed_once;

if ((ret_99FA='NO') and (ret_00SP_tmp='NO')) then ret_00SP='YES';

                              else ret_00SP=ret_00SP_tmp;

drop ret_00SP_tmp;

run;

*fa00*;

data ALL_SEM.tmp;

set ALL_SEM.uniqueID;

by id_bog;

if term='00/FA';

lvl_00FA=stu_level;

cls_00FA=class;

ret_00FA_tmp='NO'; *default=NO*;

drop stu_level class term;

run;

data ALL_SEM.listed_once;

merge ALL_SEM.listed_once ALL_SEM.tmp;

by id_bog;

run;

data ALL_SEM.listed_once;
```

```
set ALL_SEM.listed_once;
if (((ret_99FA='NO') or (ret_00SP='NO') or (ret_00SP='YES')) and
(ret_00FA_tmp='NO')) then ret_00FA='YES';
          else ret_00FA=ret_00FA_tmp;
drop ret_00FA_tmp;
run;


*Creating separate data sets for fall, fall main, spring main, spring*
data year99.y99FA;
set ALL_SEM.listed_once;
if ret_99FA='NO';
returning=0;
if ret_00FA='YES' or ret_00SP='YES' then next_returning=1;
                    else next_returning=0;
drop cls_00FA lvl_00FA ret_00FA cls_01SP lvl_01SP ret_01SP;
drop cls_01FA lvl_01FA ret_01FA cls_02SP lvl_02SP ret_02SP;
drop cls_02FA lvl_02FA ret_02FA cls_03SP lvl_03SP ret_03SP;
drop cls_03FA lvl_03FA ret_03FA cls_04SP lvl_04SP ret_04SP;
drop cls_04FA lvl_04FA ret_04FA cls_05SP lvl_05SP ret_05SP;
drop cls_05FA lvl_05FA ret_05FA;
run;
```

```
data year99.y99SP;
set ALL_SEM.listed_once;
if ((ret_99FA=") and (ret_00SP='NO'));
returning=0;
if ret_00FA='YES' or ret_00SP='YES' then next_returning=1;
                              else next_returning=0;
drop cls_00FA lvl_00FA ret_00FA cls_01SP lvl_01SP ret_01SP;
drop cls_01FA lvl_01FA ret_01FA cls_02SP lvl_02SP ret_02SP;
drop cls_02FA lvl_02FA ret_02FA cls_03SP lvl_03SP ret_03SP;
drop cls_03FA lvl_03FA ret_03FA cls_04SP lvl_04SP ret_04SP;
drop cls_04FA lvl_04FA ret_04FA cls_05SP lvl_05SP ret_05SP;
drop cls_05FA lvl_05FA ret_05FA;
run;


data year99.y99r;
set ALL_SEM.listed_once;
if ((ret_99FA='NO') and (ret_00SP='YES'));
returning=1;
if ret_00FA='YES' or ret_00SP='YES' then next_returning=1;
                              else next_returning=0;
drop cls_00FA lvl_00FA ret_00FA cls_01SP lvl_01SP ret_01SP;
drop cls_01FA lvl_01FA ret_01FA cls_02SP lvl_02SP ret_02SP;
drop cls_02FA lvl_02FA ret_02FA cls_03SP lvl_03SP ret_03SP;

drop cls_03FA lvl_03FA ret_03FA cls_04SP lvl_04SP ret_04SP;

drop cls_04FA lvl_04FA ret_04FA cls_05SP lvl_05SP ret_05SP;

drop cls_05FA lvl_05FA ret_05FA;

run;


data year99.tmp;

set  year99.y99SP;

run;

proc append base= year99.tmp data= year99.y99r;

run;

proc sort data=year99.tmp out=year99.y99SPm;

      by id_bog returning;

run;

data year99.y99SPm;

set year99.y99SPm;

by id_bog;

if first.id_bog;

run;


data year99.tmp;

set  year99.y99FA;
```

```
run;

proc append base= year99.tmp data= year99.y99r;

run;

proc sort data=year99.tmp out=year99.y99FAm;

      by id_bog returning;

run;

data year99.y99FAm;

set year99.y99FAm;

by id_bog;

if first.id_bog;

run;


proc datasets library=year99;

  delete tmp;

run;


*Combining fall semester for all years*

data ALL_SEM.ALL_FALL;

set year99.Y99FA;

YEAR='99';

LVL=lvl_99FA;

CLS=cls_99FA;
```

```
drop lvl_99FA cls_99FA ret_99FA lvl_00SP cls_00SP ret_00SP;
run;
data ALL_SEM.tmp;
set year00.Y00FA;
YEAR='00';
LVL=lvl_00FA;
CLS=cls_00FA;
drop lvl_00FA cls_00FA ret_00FA lvl_01SP cls_01SP ret_01SP;
run;
proc append base= ALL_SEM.ALL_FALL data= ALL_SEM.tmp;
run;
data ALL_SEM.tmp;
set year01.Y01FA;
YEAR='01';
LVL=lvl_01FA;
CLS=cls_01FA;
drop lvl_01FA cls_01FA ret_01FA lvl_02SP cls_02SP ret_02SP;
run;
proc append base= ALL_SEM.ALL_FALL data= ALL_SEM.tmp;
run;
data ALL_SEM.tmp;
set year02.Y02FA;
```

```
YEAR='02';
LVL=lvl_02FA;
CLS=cls_02FA;
drop lvl_02FA cls_02FA ret_02FA lvl_03SP cls_03SP ret_03SP;
run;
proc append base= ALL_SEM.ALL_FALL data= ALL_SEM.tmp;
run;
data ALL_SEM.tmp;
set year03.Y03FA;
YEAR='03';
LVL=lvl_03FA;
CLS=cls_03FA;
drop lvl_03FA cls_03FA ret_03FA lvl_04SP cls_04SP ret_04SP;
run;
proc append base= ALL_SEM.ALL_FALL data= ALL_SEM.tmp;
run;
data ALL_SEM.tmp;
set year04.Y04FA;
YEAR='04';
LVL=lvl_04FA;
CLS=cls_04FA;
drop lvl_04FA cls_04FA ret_04FA lvl_05SP cls_05SP ret_05SP;
```

```
run;

proc append base= ALL_SEM.ALL_FALL data= ALL_SEM.tmp;

run;

proc datasets library=ALL_SEM;

  delete tmp;

run;


*Create returning semester for all years*

data ALL_SEM.ALL_RETURNING;

set year99.Y99r;

YEAR='99';

LVL=lvl_99FA;

CLS=cls_99FA;

drop lvl_99FA cls_99FA ret_99FA lvl_00SP cls_00SP ret_00SP;

run;

data ALL_SEM.tmp;

set year00.Y00r;

YEAR='00';

LVL=lvl_00FA;

CLS=cls_00FA;

drop lvl_00FA cls_00FA ret_00FA lvl_01SP cls_01SP ret_01SP;

run;
```

```
proc append base= ALL_SEM.RETURNING data= ALL_SEM.tmp;
run;
data ALL_SEM.tmp;
set year01.Y01r;
YEAR='01';
LVL=lvl_01FA;
CLS=cls_01FA;
drop lvl_01FA cls_01FA ret_01FA lvl_02SP cls_02SP ret_02SP;
run;
proc append base= ALL_SEM.RETURNING data= ALL_SEM.tmp;
run;
data ALL_SEM.tmp;
set year02.Y02r;
YEAR='02';
LVL=lvl_02FA;
CLS=cls_02FA;
drop lvl_02FA cls_02FA ret_02FA lvl_03SP cls_03SP ret_03SP;
run;
proc append base= ALL_SEM.ALL_RETURNING data= ALL_SEM.tmp;
run;
data ALL_SEM.tmp;
set year03.Y03r;
```

```
YEAR='03';
LVL=lvl_03FA;
CLS=cls_03FA;
drop lvl_03FA cls_03FA ret_03FA lvl_04SP cls_04SP ret_04SP;
run;
proc append base= ALL_SEM.ALL_RETURNING data= ALL_SEM.tmp;
run;
data ALL_SEM.tmp;
set year04.Y04r;
YEAR='04';
LVL=lvl_04FA;
CLS=cls_04FA;
drop lvl_04FA cls_04FA ret_04FA lvl_05SP cls_05SP ret_05SP;
run;
proc append base= ALL_SEM.ALL_RETURNING data= ALL_SEM.tmp;
run;
proc datasets library=ALL_SEM;
  delete tmp;
run;
```

```
*Create independent variables age value, cumulative gpa, gpa value*
data fall.tmp1;
set fall.tmp1;
   ageval=2000-year(dob);
   if ageval<25 then AGE=1;
   if ((ageval>=25) and (ageval<45)) then AGE=2;
   if ageval>=45 then AGE=3;
   if ageval="" then AGE=0; **missing;
***;
   if gmat<>"" then GPAval=gmat;
            else if ((gre_q<>"") and (gre_v<>"")) then GPAval=(gre_q+gre_v)/2;
                              else if ((sat_v<>"") and (sat_m<>""))
then GPAval=(sat_v+sat_m)/2;
                                   else GPA_entry=0; **missing;
   if GPAval<450 then GPA_entry=1;  **low;
   if GPAval>=450 then GPA_entry=2; **high;
   if GPAval="" then GPA_entry=0; **missing;
***;
   if cum_gpa>=3 then GPA=2;                 **high;
   if ((cum_gpa>=2) and (cum_gpa<3)) then GPA=1;    **low;
   if cum_gpa="" then GPA=0;                 **missing;
run;
```

```
*Running frequency procedure*

proc freq data=ALL_SEM.current;

      tables cls*year /chisq expected cellchil2 norow nocol;

      weight NEXT_RETURNING;

output out=ALL_sem.tmp pchi; *pchi lrchi n nmis;

run;
```

APPENDIX C

SAS Output Tables for Different Partitioning Methods

SAS Output Tables

First Partitioning Method: $(A_1 \cup A_2 \cup A_3)$ *Tagged by Fall*

Table 21

Coefficient Estimates for the Model of Returning Students

Analysis of Maximum Likelihood Estimates					
			Standard	Wald	
Parameter	DF	Estimate	Error	Chi-Square	Pr > Chi-Square
Intercept	1	-0.3562	1.5641	0.0519	0.8198
E1 (white, non-hispanic)	1	-0.351	0.3399	1.0664	0.3018
E2 (black, non-hispanic)	1	-0.285	0.3417	0.6956	0.4043
E3 (hispanic)	1	-0.1742	0.344	0.2565	0.6125
E4 (asian, amer./pac. isl.)	1	-0.3186	0.3494	0.8312	0.3619
E5 (Amer. Indian/alaskan)	1	-0.2639	0.4073	0.4198	0.517
E6 (international)	1	-0.106	0.3549	0.0893	0.7651
E7 (other)	0	0	.	.	.
C1 (Doctorate Degree)	1	-0.4094	1.6955	0.0583	0.8092
C2 (Freshman)	1	-0.3024	0.0849	12.686	0.0004
C3 (Junior)	1	0.3457	0.0872	15.7143	<.0001
C4 (Master)	1	-0.7094	1.5199	0.2178	0.6407
C5 (Post Baccalaureate)	1	-0.7773	0.5504	1.9945	0.1579
C6 (Sophomore)	0	0	.	.	.
C7 (Senior)	1	-1.1282	0.0763	218.6753	<.0001
C8 (Sophomore)	0	0	.	.	.
L1 (Undecided)	1	-0.1685	1.6219	0.0108	0.9173
L2 (Undergraduate)	1	1.2675	1.5263	0.6896	0.4063
L3 (Post Baccalaureate)	1	1.4166	1.4226	0.9915	0.3194
L4 (Graduate Nursing)	1	-0.00973	0.1705	0.0033	0.9545
L5 (Graduate Master)	1	1.4663	0.1576	86.5258	<.0001
L6 (Graduate Doctorate)	1	1.6821	0.7666	4.8149	0.0282
L7 (Graduate Certificate)	0	0	.	.	.

Table 22

Odds Ratios for the Model of Returning Students

Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
E1 (white, non-hispanic)	0.704	0.362	1.371
E2 (black, non-hispanic)	0.752	0.385	1.469
E3 (hispanic)	0.840	0.428	1.649
E4 (asian, amer./pac. isl.)	0.727	0.367	1.442
E5 (Amer. Indian/alaskan)	0.768	0.346	1.706
E6 (international)	0.899	0.449	1.803
C1 (Doctorate Degree)	0.664	0.024	18.427
C2 (Freshman)	0.739	0.626	0.873
C3 (Junior)	1.413	1.191	1.676
C4 (Master)	0.492	0.025	9.675
C5 (Post Baccalaureate)	0.460	0.156	1.352
C7 (Senior)	0.324	0.279	0.376
L1 (Undecided)	0.845	0.035	20.296
L2 (Undergraduate)	3.552	0.178	70.742
L3 (Post Baccalaureate)	4.123	0.254	67.018
L4 (Graduate Nursing)	0.990	0.709	1.383
L5 (Graduate Master)	4.333	3.182	5.902
L6 (Graduate Doctorate)	5.377	1.197	24.159

Table 23

Partition for the Hosmer and Lemeshow Test

Partition for the Hosmer and Lemeshow Test					
		next_returning = 1		next_returning = 0	
Group	Total	Observed	Expected	Observed	Expected
1	1227	247	240.44	980	986.56
2	1694	585	583.92	1109	1110.08
3	1187	480	486.74	707	700.26
4	94	47	46.97	47	47.03
5	2299	1161	1178.07	1138	1120.93
6	1044	574	560.39	470	483.61
7	1220	689	695.74	531	524.26
8	1371	875	867.16	496	503.84
9	2081	1461	1459.57	620	621.43

Second Partitioning Method: $(A_1 \cup A_2) \cup A_3$

Table 24

Coefficient Estimates for the Model of Returning Students (Fall Main)

Analysis of Maximum Likelihood Estimates										
	FallMain									
			Std.	Wald				Std.	Wald	
Parameter	DF	Estimate	Error	Chi-Square	Pr > ChiSq	DF	Estimate	Error	Chi-Square	Pr > ChiSq
Int.	1	-0.9261	1.7165	0.2911	0.5895	1	-1.2973	1.6298	0.6336	0.426
E1 (white, non-hispanic)	1	0.0867	0.3396	0.0651	0.7986	1	0.3988	0.5259	0.575	0.4483
E2 (black, non-hispanic)	1	0.0965	0.3419	0.0797	0.7777	1	0.5675	0.5458	1.0813	0.2984
E3 (hispanic)	1	0.2407	0.3449	0.4871	0.4852	1	0.3811	0.5706	0.4461	0.5042
E4 (asian, amer./pac. isl.)	1	0.0878	0.3521	0.0621	0.8032	1	0.8082	0.5855	1.9056	0.1675
E5 (Amer. Indian/alaskan)	1	0.0652	0.42	0.0241	0.8766	1	-11.5675	378.7	0.0009	0.9756
E6 (international)	1	0.2907	0.3585	0.6575	0.4174	0	0	.	.	.
E7 (other)	0	0	.	.	.	0	0	.	.	.
C1 (Doctorate Degree)	1	0.2153	1.8361	0.0138	0.9066	1	-0.7167	1.5668	0.2093	0.6474
C2 (Freshman)	1	-0.4205	0.0994	17.8901	<.0001	1	-0.1567	0.3621	0.1872	0.6652
C3 (Junior)	1	0.2293	0.1032	4.9394	0.0263	1	0.5327	0.3395	2.4615	0.1167
C4 (Master)	1	-0.0364	1.6747	0.0005	0.9826	1	-2.0882	1.6143	1.6733	0.1958
C5 (Post Baccalaureate)	1	0.3044	0.8926	0.1163	0.7331	1	-0.9975	1.5442	0.4173	0.5183
C6 (Sophomore)	0	0	.	.	.	0	0	.	.	.
C7 (Senior)	1	-1.1952	0.0898	177.0651	<.0001	1	-0.2052	0.2951	0.4833	0.4869
C8 (Sophomore)	0	0	.	.	.	0	0	.	.	.
L1 (Undecided)	1	0.018	1.7433	0.0001	0.9918	0	0	.	.	.
L2 (Undergraduate)	1	1.9091	1.6818	1.2886	0.2563	1	-1.3572	1.5267	0.7903	0.374
L3 (Post Baccalaureate)	1	0.8759	1.4248	0.3779	0.5387	0	0	.	.	.
L4 (Graduate Nursing)	1	0.0323	0.1922	0.0282	0.8667	1	0.1752	0.4973	0.1241	0.7246

Table 24 (continued)

Analysis of Maximum Likelihood Estimates										
	FallMain									
Parameter	DF	Estimate	Std. Error	Wald Chi-Square	Pr > ChiSq	DF	Estimate	Std. Error	Wald Chi-Square	Pr > ChiSq
L5 (Graduate Master)	1	1.2711	0.1769	51.6608	<.0001	1	0.733	0.4808	2.3243	0.1274
L6 (Graduate Doctorate)	1	1.6227	0.7724	4.4138	0.0356	0	0	.	.	.
L7 (Graduate Certificate)	0	0	.	.	.	0	0	.	.	.

Table 25

Odds Ratios for the Model of Returning Students (Fall Main)

Odds Ratio Estimates						
	FallMain			Spring Subset		
		95% Wald			95% Wald	
Effect	Point Estimate	Confidence Limits		Point Estimate	Confidence Limits	
E1 (white, non-hispanic)	1.091	0.56	2.122	1.49	0.532	4.177
E2 (black, non-hispanic)	1.101	0.564	2.153	1.764	0.605	5.141
E3 (hispanic)	1.272	0.647	2.501	1.464	0.478	4.48
E4 (asian, amer./pac. isl.)	1.092	0.547	2.177	2.244	0.712	7.07
E5 (Amer. Indian/alaskan)	1.067	0.469	2.431	<0.001	<0.001	>999.999
E6 (international)	1.337	0.662	2.701	.	.	.
C1 (Doctorate Degree)	1.24	0.034	45.329	0.488	0.023	10.529
C2 (Freshman)	0.657	0.54	0.798	0.855	0.42	1.739
C3 (Junior)	1.258	1.027	1.539	1.703	0.876	3.314
C4 (Master)	0.964	0.036	25.685	0.124	0.005	2.932
C5 (Post Baccalaureate)	1.356	0.236	7.797	0.369	0.018	7.607
C7 (Senior)	0.303	0.254	0.361	0.814	0.457	1.452
L1 (Undecided)	1.018	0.033	31.023	.	.	.
L2 (Undergraduate)	6.747	0.25	182.267	0.257	0.013	5.129
L3 (Post Baccalaureate)	2.401	0.147	39.194	.	.	.
L4 (Graduate Nursing)	1.033	0.709	1.505	1.191	0.45	3.158
L5 (Graduate Master)	3.565	2.521	5.042	2.081	0.811	5.34
L6 (Graduate Doctorate)	5.067	1.115	23.023	.	.	.

Table 26

Partition for the Hosmer and Lemeshow Test (Fall Main)

Partition for the Hosmer and Lemeshow Test										
	FallMain					Spring Subset				
		next_returning = 1		next_returning = 0			next_returning = 1		next_returning = 0	
Group	Total	Observed	Expected	Observed	Expected	Total	Observed	Expected	Observed	Expected
1	863	267	265.09	596	597.91	197	9	9.26	188	187.74
2	1086	514	508.96	572	577.04	346	24	19.62	322	326.38
3	955	487	493.87	468	461.13	160	8	11.34	152	148.66
4	1896	1125	1132.96	771	763.04	394	29	30.99	365	363.01
5	1037	646	639.48	391	397.52	269	24	23.25	245	245.75
6	895	584	588.58	311	306.42	163	14	15.36	149	147.64
7	757	550	546.08	207	210.92	403	36	38.3	367	364.7
8	974	734	730.59	240	243.41	249	33	28.24	216	220.76
9	1247	983	984.39	264	262.61	326	52	52.63	274	273.37

Table 27

Coefficient Estimates for the Model of Returning Students (Spring Subset)

Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-1.2973	1.6298	0.6336	0.4260
E1 (white, non-hispanic)	1	0.3988	0.5259	0.5750	0.4483
E2 (black, non-hispanic)	1	0.5675	0.5458	1.0813	0.2984
E3 (hispanic)	1	0.3811	0.5706	0.4461	0.5042
E4 (asian, amer./pac. isl.)	1	0.8082	0.5855	1.9056	0.1675
E5 (Amer. Indian/alaskan)	1	-11.5675	378.7	0.0009	0.9756
E6 (international)	0	0	.	.	.
E7 (other)	0	0	.	.	.
C1 (Doctorate Degree)	1	-0.7167	1.5668	0.2093	0.6474
C2 (Freshman)	1	-0.1567	0.3621	0.1872	0.6652
C3 (Junior)	1	0.5327	0.3395	2.4615	0.1167
C4 (Master)	1	-2.0882	1.6143	1.6733	0.1958
C5 (Post Baccalaureate)	1	-0.9975	1.5442	0.4173	0.5183
C6 (Sophomore)	0	0	.	.	.
C7 (Senior)	1	-0.2052	0.2951	0.4833	0.4869
C8 (Sophomore)	0	0	.	.	.
L1 (Undecided)	0	0	.	.	.
L2 (Undergraduate)	1	-1.3572	1.5267	0.7903	0.3740
L3 (Post Baccalaureate)	0	0	.	.	.
L4 (Graduate Nursing)	1	0.1752	0.4973	0.1241	0.7246
L5 (Graduate Master)	1	0.7330	0.4808	2.3243	0.1274
L6 (Graduate Doctorate)	0	0	.	.	.
L7 (Graduate Certificate)	0	0	.	.	.

Table 28

Odds Ratios for the Model of Returning Students (Spring Subset)

Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
E1 (white, non-hispanic)	3.177	0.431	23.398
E2 (black, non-hispanic)	4.210	0.550	32.242
E3 (hispanic)	3.414	0.424	27.492
E4 (asian, amer./pac. isl.)	6.807	0.841	55.060
E5 (Amer. Indian/alaskan)	<0.001	<0.001	>999.999
C1 (Doctorate Degree)	0.329	<0.001	326.389
C2 (Freshman)	0.098	0.012	0.813
C3 (Junior)	2.206	0.906	5.374
C4 (Master)	0.021	<0.001	26.309
C5 (Post Baccalaureate)	0.118	<0.001	115.107
C7 (Senior)	0.682	0.294	1.581
L2 (Undergraduate)	0.096	<0.001	89.833
L4 (Graduate Nursing)	2.433	0.313	18.944
L5 (Graduate Master)	7.140	0.967	52.689

Table 29

Partition for the Hosmer and Lemeshow Test (Spring Subset)

Partition for the Hosmer and Lemeshow Test					
		next_returning = 1		next_returning = 0	
Group	Total	Observed	Expected	Observed	Expected
1	222	1	1.06	221	220.94
2	142	2	1.72	140	140.28
3	346	9	7.76	337	338.24
4	57	2	1.37	55	55.63
5	394	10	11.48	384	382.52
6	245	10	8.36	235	236.64
7	243	11	11.54	232	231.46
8	434	26	27.28	408	406.72
9	269	20	23.19	249	245.81
10	155	24	21.24	131	133.76

Third Partitioning Method: $A_1 \cup (A_2 \cup A_3)$

Table 30

Coefficient Estimates for the Model of Returning Students (Fall Subset)

Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	9.5025	177.1	0.0029	0.9572
E1 (white, non-hispanic)	1	-0.0161	0.3379	0.0023	0.9619
E2 (black, non-hispanic)	1	-0.0108	0.3446	0.0010	0.9749
E3 (hispanic)	1	0.1414	0.3534	0.1601	0.6890
E4 (asian, amer./pac. isl.)	1	-0.0230	0.3756	0.0038	0.9511
E5 (Amer. Indian/alaskan)	1	-0.1474	0.5457	0.0730	0.7871
E6 (international)	1	0.0865	0.3824	0.0512	0.8210
E7 (other)	0	0	.	.	.
C1 (Doctorate Degree)	1	-10.3026	177.1	0.0034	0.9536
C2 (Freshman)	1	-0.2135	0.1339	2.5422	0.1108
C3 (Junior)	1	0.1271	0.1511	0.7079	0.4001
C4 (Master)	1	-10.4818	177.1	0.0035	0.9528
C5 (Post Baccalaureate)	1	0.5814	0.9058	0.4120	0.5210
C6 (Sophomore)	0	0	.	.	.
C7 (Senior)	1	-0.5463	0.1931	8.0036	0.0047
C8 (Sophomore)	0	0	.	.	.
L1 (Undecided)	1	-10.5183	177.1	0.0035	0.9526
L2 (Undergraduate)	1	-8.6189	177.1	0.0024	0.9612
L3 (Post Baccalaureate)	1	-9.3578	177.1	0.0028	0.9579
L4 (Graduate Nursing)	1	-0.00503	0.2599	0.0004	0.9846
L5 (Graduate Master)	1	1.9722	0.2495	62.5024	<.0001

Table 30 (continued)

Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
L6 (Graduate Doctorate)	1	2.0018	0.8797	5.1784	0.0229
L7 (Graduate Certificate)	0	0	.	.	.

Table 31

Odds Ratios for the Model of Returning Students (Fall Subset)

Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
E1 (white, non-hispanic)	0.984	0.507	1.908
E2 (black, non-hispanic)	0.989	0.503	1.944
E3 (hispanic)	1.152	0.576	2.303
E4 (asian, amer./pac. isl.)	0.977	0.468	2.040
E5 (Amer. Indian/alaskan)	0.863	0.296	2.515
E6 (international)	1.090	0.515	2.307
C1 (Doctorate Degree)	<0.001	<0.001	>999.999
C2 (Freshman)	0.808	0.621	1.050
C3 (Junior)	1.136	0.845	1.527
C4 (Master)	<0.001	<0.001	>999.999
C5 (Post Baccalaureate)	1.789	0.303	10.557
C7 (Senior)	0.579	0.397	0.845
L1 (Undecided)	<0.001	<0.001	>999.999
L2 (Undergraduate)	<0.001	<0.001	>999.999
L3 (Post Baccalaureate)	<0.001	<0.001	>999.999
L4 (Graduate Nursing)	0.995	0.598	1.656
L5 (Graduate Master)	7.187	4.407	11.718
L6 (Graduate Doctorate)	7.402	1.320	41.511

Table 32

Odds Ratios for the Model of Returning Students (Spring Main)

Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
E1 (white, non-hispanic)	1.490	0.532	4.177
E2 (black, non-hispanic)	1.764	0.605	5.141
E3 (hispanic)	1.464	0.478	4.480
E4 (asian, amer./pac. isl.)	2.244	0.712	7.070
E5 (Amer. Indian/alaskan)	<0.001	<0.001	>999.999
C1 (Doctorate Degree)	0.488	0.023	10.529
C2 (Freshman)	0.855	0.420	1.739
C3 (Junior)	1.703	0.876	3.314
C4 (Master)	0.124	0.005	2.932
C5 (Post Baccalaureate)	0.369	0.018	7.607
C7 (Senior)	0.814	0.457	1.452
L2 (Undergraduate)	0.257	0.013	5.129
L4 (Graduate Nursing)	1.191	0.450	3.158
L5 (Graduate Master)	2.081	0.811	5.340

Table 33

Partition for the Hosmer and Lemeshow Test (Spring Main)

Partition for the Hosmer and Lemeshow Test					
		next_returning = 1		next_returning = 0	
Group	Total	Observed	Expected	Observed	Expected
1	377	98	100.23	279	276.77
2	382	151	150.90	231	231.10
3	486	312	318.71	174	167.29
4	377	252	250.49	125	126.51
5	401	288	280.75	113	120.25
6	149	111	106.40	38	42.60
7	740	540	537.57	200	202.43
8	478	351	348.52	127	129.48
9	471	345	354.43	126	116.57

Table 34

Coefficient Estimates for the Model of Returning Students (Fall Main)

Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-0.5396	1.4376	0.1409	0.7074
E1 (white, non-hispanic)	1	-0.2735	0.1351	4.0976	0.0429
E2 (black, non-hispanic)	1	-0.2078	0.1440	2.0833	0.1489
E3 (hispanic)	1	-0.0354	0.1500	0.0557	0.8134
E4 (asian, amer./pac. isl.)	1	-0.1125	0.1653	0.4631	0.4962
E5 (Amer. Indian/alaskan)	1	0.1280	0.3002	0.1817	0.6699
E6 (international)	0	0	.	.	.
E7 (other)	0	0	.	.	.
C1 (Doctorate Degree)	1	10.8663	159.5	0.0046	0.9457
C2 (Freshman)	1	-0.5278	0.1054	25.0648	<.0001
C3 (Junior)	1	-0.3430	0.0972	12.4459	0.0004
C4 (Master)	1	-0.8756	1.4178	0.3814	0.5368
C5 (Post Baccalaureate)	1	-0.6040	0.4441	1.8495	0.1738
C6 (Sophomore)	0	0	.	.	.
C7 (Senior)	1	-1.7012	0.0929	335.0113	<.0001
C8 (Sophomore)	0	0	.	.	.
L1 (Undecided)	1	-10.6609	275.3	0.0015	0.9691
L2 (Undergraduate)	1	1.4506	1.4302	1.0287	0.3105
L3 (Post Baccalaureate)	1	0.9302	1.5003	0.3844	0.5352
L4 (Graduate Nursing)	1	0.3201	0.2214	2.0904	0.1482
L5 (Graduate Master)	1	1.1900	0.2080	32.7415	<.0001
L6 (Graduate Doctorate)	1	-9.6709	159.5	0.0037	0.9517
L7 (Graduate Certificate)	0	0	.	.	.

Table 35

Odds Ratios for the Model of Returning Students (Fall Main)

Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
E1 (white, non-hispanic)	0.761	0.584	0.991
E2 (black, non-hispanic)	0.812	0.613	1.077
E3 (hispanic)	0.965	0.719	1.295
E4 (asian, amer./pac. isl.)	0.894	0.646	1.236
E5 (Amer. Indian/alaskan)	1.137	0.631	2.047
C1 (Doctorate Degree)	>999.999	<0.001	>999.999
C2 (Freshman)	0.590	0.480	0.725
C3 (Junior)	0.710	0.587	0.859
C4 (Master)	0.417	0.026	6.707
C5 (Post Baccalaureate)	0.547	0.229	1.305
C7 (Senior)	0.182	0.152	0.219
L1 (Undecided)	<0.001	<0.001	>999.999
L2 (Undergraduate)	4.266	0.259	70.371
L3 (Post Baccalaureate)	2.535	0.134	47.982
L4 (Graduate Nursing)	1.377	0.892	2.126
L5 (Graduate Master)	3.287	2.187	4.942
L6 (Graduate Doctorate)	<0.001	<0.001	>999.999

Table 36

Partition for the Hosmer and Lemeshow Test (Fall Main)

Partition for the Hosmer and Lemeshow Test					
		next_returning = 1		next_returning = 0	
Group	Total	Observed	Expected	Observed	Expected
1	852	166	167.81	686	684.19
2	1160	298	296.54	862	863.46
3	659	189	188.65	470	470.35
4	1433	528	541.46	905	891.54
5	875	379	356.35	496	518.65
6	782	390	407.05	392	374.95
7	748	441	429.89	307	318.11
8	800	475	476.16	325	323.84
9	1047	691	693.08	356	353.92

Forth Partitioning Method: $A_1 \cup A_2 \cup A_3$ (A_2 tagged by Spring)

Table 37

Coefficient Estimates for the Model of Returning Students (Intersection Subset)

Analysis of Maximum Likelihood Estimates					
			Standard	Wald	
Parameter	DF	Estimate	Error	Chi-Square	Pr > Chi-Square
Intercept	1	1.5018	1.448	1.0756	0.2997
E1 (white, non-hispanic)	1	-0.2656	0.1604	2.7433	0.0977
E2 (black, non-hispanic)	1	-0.2723	0.1709	2.5395	0.111
E3 (hispanic)	1	-0.0656	0.1781	0.1356	0.7127
E4 (asian, amer./pac. isl.)	1	-0.1722	0.196	0.7713	0.3798
E5 (Amer. Indian/alaskan)	1	-0.0379	0.3478	0.0119	0.9132
E6 (international)	0	0	.	.	.
E7 (other)	0	0	.	.	.
C1 (Doctorate Degree)	1	10.4751	263.8	0.0016	0.9683
C2 (Freshman)	1	-0.6309	0.1448	18.9923	<.0001
C3 (Junior)	1	-0.9457	0.1282	54.4434	<.0001
C4 (Master)	1	-1.7988	1.4206	1.6033	0.2054
C5 (Post Baccalaureate)	1	-0.6952	0.6016	1.3353	0.2479
C6 (Sophomore)	0	0	.	.	.
C7 (Senior)	1	-2.0231	0.1244	264.631	<.0001
C8 (Sophomore)	0	0	.	.	.
L1 (Undecided)	1	-13.698	459.8	0.0009	0.9762
L2 (Undergraduate)	1	0.3626	1.4362	0.0637	0.8007
L3 (Post Baccalaureate)	1	-0.3572	1.5606	0.0524	0.8189
L4 (Graduate Nursing)	1	0.7002	0.2675	6.8532	0.0088
L5 (Graduate Master)	1	0.5379	0.2442	4.8504	0.0276
L6 (Graduate Doctorate)	1	-10.791	263.8	0.0017	0.9674
L7 (Graduate Certificate)	0	0	.	.	.

Table 38

Odds Ratios for the Model of Returning Students (Intersection Subset)

Odds Ratio Estimates			
		95% Wald	
Effect	Point Estimate	Confidence Limits	
E1 (white, non-hispanic)	0.767	0.560	1.050
E2 (black, non-hispanic)	0.762	0.545	1.065
E3 (hispanic)	0.937	0.661	1.328
E4 (asian, amer./pac. isl.)	0.842	0.573	1.236
E5 (Amer. Indian/alaskan)	0.963	0.487	1.903
C1 (Doctorate Degree)	>999.999	<0.001	>999.999
C2 (Freshman)	0.532	0.401	0.707
C3 (Junior)	0.388	0.302	0.499
C4 (Master)	0.166	0.010	2.679
C5 (Post Baccalaureate)	0.499	0.153	1.622
C7 (Senior)	0.132	0.104	0.169
L1 (Undecided)	<0.001	<0.001	>999.999
L2 (Undergraduate)	1.437	0.086	23.987
L3 (Post Baccalaureate)	0.700	0.033	14.903
L4 (Graduate Nursing)	2.014	1.192	3.402
L5 (Graduate Master)	1.712	1.061	2.764
L6 (Graduate Doctorate)	<0.001	<0.001	>999.999

Table 39

Partition for the Hosmer and Lemeshow Test (Intersection Subset)

Partition for the Hosmer and Lemeshow Test					
		next_returning = 1		next_returning = 0	
Group	Total	Observed	Expected	Observed	Expected
1	270	105	104.12	165	165.88
2	666	263	263.39	403	402.61
3	419	198	193.99	221	225.01
4	1030	502	508.64	528	521.36
5	578	312	309.03	266	268.97
6	479	291	287.73	188	191.27
7	570	375	374.89	195	195.11
8	565	398	401.13	167	163.87
9	599	437	438.60	162	160.40
10	673	561	560.49	112	112.51