

https://dspace.org/

Analysing first year students' performance in the Commerce Faculty at the University of the Witwatersrand

Yathavan, Vasuki

2009-09-16T07:10:58Z
http://hdl.handle.net/10539/7268

Downloaded from a DSpace repository. https://dspace.org/

CHAPTER 1

INTRODUCTION

The intricacies of coping at University- the first-year experience- are the focus of ongoing study by the higher education sector internationally.

Students are selected mainly on their matriculation examination results in the Faculty of Commerce at the University of the Witwatersrand. In other research in South Africa, only a very small percentage of those students with a school result average of below 70% obtained a first-year University average performance of 50% or better (Roux, 2004). According to the report of the working group on retention and throughput at the University of the Witwatersrand, of the students who originally registered in the 1992 to 1998 cohorts less than 50% on average graduate in most of the degrees; and between 15% and 20 % of students drop out. In the Commerce Faculty 24%, 32% and 31% of students had dropped out in the years 2000, 2001 and 2002 respectively during the first year; and a further 9%, 9% and 12% respectively are excluded (Oracle System, Management Information Unit at University of the Witwatersrand, 2007).

Exclusion rates for first year students in the Faculty of Commerce were 38%, 40% and 38% in the years 2000, 2001 and 2002 respectively; A further 15%, 16% and 12%, respectively have been required to repeat the first year (Oracle system by Management Information Unit at University of the Witwatersrand, 2007). The problem of this data statement is that an average of only 47% of first year students has passed into their second year.

Students entering South African Universities come from a wide range of social and cultural backgrounds that give them very different life experiences, different educational opportunities and a great variety of expectations, needs and academic potential (Chickte & Brand, 1996; Goduka, 1996). This situation also occurs in other countries that have shifted the focus of higher education from elitism to mass opportunities (McKenzie & Schweitzer, 2001). When students are admitted to a higher education institution there is a tacit assumption that they will be capable of successfully completing the course in which they are permitted to enrol. To knowingly admit students who, for whatever reason, have no chances of academic success would be immoral.

Despite the existence of an extensive literature (e.g. Aitken, 1982; Bargate, 1999) on the factors that are associated with University students' performance, the need remains for the development of comprehensive models of University student adjustment and retention that not only captures the underlying structural relationships of the process, but that is also operational in the sense that it can be effectively used by the individual institution. The theories formulated by both Spady (1970) and Tinto (1975) identify several independent variables that affect student retention, with only a few major variables having a direct effect on retention and the remaining variables affecting retention indirectly.

Academic development programmes have been in place at the University of the Witwatersrand for more than 20 years and the proposed solutions to the current unsatisfactory throughput rate should not be assumed to be as simple as doing more of the same or even doing the same things in slightly different ways (Success at the University of the Witwatersrand, Strategic Planning Division, University of the Witwatersrand, 2003). The University of the Witwatersrand needs to achieve improvements and find ways of doing this. In recent years, the retention and throughput of students have assumed renewed importance in the higher education sector (Ian, Nan & Jane, 2007). There have been increased pressures from the government and its agencies to demonstrate value for money in the use of public finance.

For most of the undergraduate degrees there are statistically significant differences in performance between the successes achieved by different gender and race groups. Black students have historically done worse at this University than have White students and women have fared better than men. As proportionally more Black students face economic and social disadvantage than do White students, there are other national issues which impact on students' ability to study successfully (Success at the University of the Witwatersrand, strategic planning division, University of the Witwatersrand, 2003). The University has taken the approach that preparedness and other factors also affect performance. Importantly, matriculation aggregate and some of the more common matriculation courses will be considered in this report.

Historically for the B.Com (ordinary) proportionally more Whites drop out than do Africans and Coloureds (Success at the University of the Witwatersrand, strategic planning division, University of the Witwatersrand, 2003). Race will not be considered as the predictor variable on students' performance in this report. Undoubtedly there is a need for improvements in teaching at many schools in the country and some young people come to University with less of the fundamental learning in place than do others. The University is not in a position to address longstanding legacies of inadequate schooling on its own.

This chapter will now present a brief summary of the literature reviewed in chapter two including the theoretical framework guiding this study and the research regarding the first-year student academic performance. The chapter will continue with the purpose of the study, the research questions and the research methodology.

1.1 Literature review

This section contains a theoretical discussion of the statistical methods used in the analysis: Chi-squared Automatic Interaction Detection (CHAID) analysis and Multinomial Logistic Regression.

Whenever we measure a variable it has to be on some type of scale. The following scales are delivered in order of increasing complexity (Agresti, 1990).

Nominal scales- These are not really values at all, but are instead numbers used to differentiate objects.

Ordinal scales- Ordinal scales use numbers to put objects in order.

Interval scales-Interval scales contain an ordinal scale, but have the added feature that the distance between scale units is always the same.

CHAID (Kass, 1980) is the acronym for *Chi*-squared Automatic Interaction Detector. The "*Chi-Squared*" part of the name arises because the technique essentially involves automatically constructing many cross-tabulations, and working out statistical significance of the proportions (Hoare, 2004). The most significant relationships are used to control the structure of a tree diagram. It is an exploratory method used to study the relationship between a dependent variable and a series of predictor variables. CHAID modelling selects a set of predictors and their interactions that optimally predict the dependent measure.

Logistic regression is a type of predictive model that can be used when the dependent variable is a categorical variable (Cox, 1970). It does not involve decision trees and is more akin to nonlinear regression. Binary logistic regression is the two-group logistic regression model. Multinomial Logistic Regression is the extension for the (binary) logistic regression when the categorical dependent variable has more than two levels. It can be divided into two cases: ordinal response and nominal response.

It has been known for a long time that the canonical parameter for the binomial distribution is obtained by a logistic transformation of the probability parameter; (for example Lehmann (1959) or Barndorff-Nielsen (1978)). The conditions for existence of unique solutions to the likelihood equations were given by Albert and Andersson (1984).

When the class dependent variable takes on more than two outcomes or classes, the multinomial regression model, an extension of the Binomial Logistic Regression (BLR)

model, can be used to predict class membership. In a Multinomial Logistic Regression model, the estimates for the parameter can be identified compared to a baseline category (Long, 1997). In principle, the strategies and methods for multivariable modelling with a multinomial dependent variable are identical to those for the binary dependent variable.

The practice of using school matriculation results as the sole or primary determinant for University entrance is common in many countries, such as Australia and USA, where there is strong competition for University entrance (McKenzie & Schweitzer, 2001). Huysamen (1999) pointed out that the poor predictability of educationally disadvantaged students' first-year performance was a finding not unique to South Africa and discussed several psychometric explanations for this finding. According to Nunns and Ortlepp (1994) the basis of fair selection is the establishment of accurate predictors of academic performance.

Jawitz (1995) claimed that matriculation results (with the exception of the then Department of Education and Training matriculants) correlated well with success at University level, particularly at first-year level. Nobel and Sawyer (1997) showed that academic ability, as measured by high school grades has predictive validity sufficient to set admissions criteria for selection. Matriculants from educationally disadvantaged high schools increasingly succeed in narrowing the gap between their academic performance and that of their counterparts from educationally non-disadvantaged high schools (Huysamen, 2001). School achievement was the best cognitive predictor of average first-year performance.

School marks for Mathematics, Physical Science and English were all related to first year performance (Eeden, Beer & Coetzee, 2001). In the history of South Africa, language-in-education policy has always been a contentious issue (Alexander, 2001). Delvare (1995) maintains that English as medium of instruction presents a problem because most Black students have English as a second or third language.

Irrespective of their high school background, the tertiary academic-performance of women was under predicted and that of men was over predicted when students were grouped together in terms of gender (Huysamen, 2001).

It is of interest that, when the broad findings of the Department of Education's (DoE) 2000 cohort study appeared in the press, virtually all of the reported responses attributed the high attrition rates to 'money and poor schooling' (Mail & Guardian, 2006).

The purpose of this study, the research questions and the research methodology chosen to examine those questions will now be presented.

1.2 Purpose of study

It is necessary to have entry requirements that permit fair, valid and reliable student selection decisions to be made. If there are any possible predictors that influence the first year students' performance in the Faculty of Commerce at the University of the Witwatersrand, the purpose of this research is to find these important predictors. Here, theoretical models those take into account a variety of attributes and pre-university experience that may affect a student's performance in the first year would be suggested and tested. This may result in possible changes to the entry requirements in an endeavour to increase the first year pass rate.

The building of a classification model based on a three-category- completed, excluded, returned- and of students' performance will be illustrated. Completed means that the students pass into their second year; returned means that students are enrolled in the Faculty but have not achieved sufficient marks to pass to the second year of study i.e. they are required to repeat first year; excluded means that students who have cancelled their registration, or have been excluded either for financial or academic reasons. This category includes drop-outs.

The hypothesis of this report is the finding the some of the important determinants of student first year performance. This report has other possible objectives including assessment of the effect, or relationship between, explanatory variables on the response and a general description of data structure. For example, if a student has matriculation aggregate between 60% and 70%, then his/her pass chance may be 60%. But if he/she takes Accountancy as a matriculation subject his/her pass chance may be increased or decreased. If so, there should be an interaction between aggregate and Accountancy.

1.3 Research questions

The study aims to examine the academic performance of first year students of Faculty of Commerce, University of the Witwatersrand with respect to some predictor variables. Operationally, the following questions to be answered by the study were formulated:

- Do matriculation Aggregate, some common matriculation courses (Accounting, Biology, English, History, Mathematics and Physical Science), Gender, Age and previous Institution type predict first year performance?
- Which variables, at what stages, are most efficient to predict "completed"?
- How do the different factors combine and interact with each other?

1.4 Methodology

In this research report, the model will be calibrated using 2003 and 2004 Commerce first year cohort data. This data is available for about 2500 students with their pre-university information and first year performance. For the purpose of this report, students' inclusion is defined in terms of full time students only. Full time students must enrol in four courses per term, each course requiring 175 working hours per semester.

These data were collected from the Oracle Student System (OSS) by the Management Information Unit (MIU) at the University of the Witwatersrand. It provides management and statistics for the University of the Witwatersrand. OSS has been the system from 2nd

January 2007 at the University of the Witwatersrand; this data system replaced the Student Information Record System (SIRS).

The cohort data of 2005 and 2006 data have also been collected to validate the model. In addition to the students' performance, data on other variables (student number, date of birth, gender, name of school, previous institution type, results of matriculation courses, school aggregate, and degree type) are gathered for each of the student.

All students take English at higher grade as a first language or second language and Mathematics as higher grade or standard grade and these two school subjects with certain minima are the minimum requirements for the admission to the Commerce Faculty (www.wits.ac.za). History, Biology, Accountancy and Physical Science are the most popular courses that potential Commerce Faculty students take at school. Each student had a previous institution type:
School – those who came directly from school without any further or higher education;
Further Education – those who came from colleges;
Higher Education – those who came from other University or Faculties other than Commerce Faculty at the same University.

Data cleaning will be done in three parts: the first part will involve checking each entry for consistency; the second part will involve selecting variables using CHAID analysis over 2003 and 2004 data and validating using 2005 and 2006 data; the third part will involve building prediction models using Multinomial Logistic Regression .

Unnecessary variables (student number, date of birth, name of school and degree type) for the analysis will be deleted from the database. Every dataset contains some errors. Errors and inconsistencies will first be detected by checking every student's details row by row in Excel and corrected. Some students in the database do not have any useful information for analysis. For example, they don't have any matriculation course details. Such students will be deleted from the database. Duplicate information will also be

eliminated. There will be no imputation for the missing values as the data set contains only less than 1% of missing values. Age will be introduced in the database.

CHAID is a large sample procedure. An appropriate SAS dataset will be created containing student performance and the other continuous and categorical predictor variables. Two SAS datasets will be created one will be for training data set (2003 and 2004, 2513 students) to derive the model and other will be for validation (2005 and 2006, 2274 students).

SAS and R will be used in the study for both analyses and the creation of some graphs respectively.

A SAS macro is a way of defining parts of or a collection of SAS statements which can be carried out repeatedly or which can substitute names of datasets and variables for symbolic names. The CHAID macro is a SAS application for performing classification models based on decision trees. The CHAID macro generates a SAS dataset that describes a decision tree computed from a training dataset to predict a specified categorical response variable from one or more predictor variables. SAS/CORE, SAS/BASE, SAS/IML, SAS/GRAPH and optional SAS/OR must be licensed and installed at the site to run the programme (Fernandez, 2003).

R is a programming language designed for statistical analysis. It was originally created by Ross Ihaka and Robert Gentleman at the University of Auckland, New Zealand and now developed by the R development Core Team. R uses a command line interface, though several graphical user interfaces are available. R's strength is its graphical facilities, which produce publication-quality graphs which can include mathematical symbols.

In Chapter 4 the results of the CHAID and Multinomial Logistic Regression analyses as applied to Faculty of Commerce first year students' data are reported. Results for each method are presented. Prediction accuracies are also presented.

Chapter 5 discusses the results from both the CHAID and Multinomial Logistic Regression methodologies. This Chapter also makes final recommendations and conclusions to the study.

CHAPTER 2

LITERATURE REVIEW AND METHODOLOGY

This chapter contains a theoretical discussion of the statistical methods used in the analysis: Chi-squared Automatic Interaction (CHAID) analysis and Multinomial Logistic Regression for analysing first year students' performance in Faculty of Commerce, University of the Witwatersrand.

2.1 Scales of measure

Categorical variables for which levels do not have a natural ordering are called nominal variables. Examples of nominal variables are: religious affiliation (Catholic, Jewish, Protestant, other), mode of transportation (automobile, bus, subway, bicycle, other), choice of residence (house, apartment, condominium, other), race, gender, and marital status. For nominal variables, the order of listing of the categories is irrelevant in the statistical analysis (Agresti, 1990).

Many categorical variables do have ordered levels. Such variables are called ordinal variables. Examples of ordinal variables are: social class (upper, middle, lower), attitude toward legalisation of abortion (strongly disapprove, approve, strongly approve), appraisal of company's inventory level (too low, about right, too high), and diagnosis of whether patient has multiple sclerosis (certain, probable, unlikely, definitely not). Ordinal variables clearly order the categories, but absolute distances between categories are unknown. While we can conclude that a person categorized as "moderate" is more liberal

than a person categorized as “conservative,” we cannot give a numerical value for how much more liberal that person is.

An interval variable is one that does have numerical distances between any two levels of the scale. On this scale, one unit on the scale represents the same magnitude on the trait or characteristic being measured across whole range of the scale. The numbers assigned to objects have all the features of ordinal measurement, and in addition equal differences between measurements represent equivalent intervals. That is, differences between arbitrary pairs of measurements can be meaningfully compared. For example, blood pressure level, functional life length of television set, length of prison term, income, and age are all interval variables.

2.2 Statistical methods

CHAID is one of the oldest tree classification methods originally proposed by Kass (1980); according to Ripley, 1996, the CHAID algorithm is a descendent of THAID developed by Morgan & Messenger, 1973. It is a technique that detects interaction between variables. As this statistic is only approximately chi-squared distributed, a large sample size is required (Eherler & Lehmann, 2001). The dependent measure may be a qualitative (nominal or ordinal) one or a quantitative one. Like other decision trees, its advantages are that’s its output is highly visual and can be easy to interpret.

Regression informs us how one variable is related to another-or to several others (Wonnacott & Wonnacott, 1981). Logistic regression is part of a category of statistical models called the generalised linear models. Logistic regression is powerful in its ability to estimate the individual effects of continuous or categorical independent variables on categorical dependent variables (Wright, 1995). Binomial (or binary) logistic regression is a form of regression which is used when the dependent variable is a dichotomy and the independent variables are of any type. It uses Maximum-Likelihood Estimation (MLE) after converting the binary response into a logit value (the natural log of the odds of the

response occurring or not) and estimates the probability of a given event occurring. It makes no assumption about the distribution of the independent variables.

McFadden (1974) proposed a modification of the logistic regression model and called it a discrete choice model. As a result the model is frequently referred to as the discrete choice model in business and econometric literature while it is called the multinomial, polychotomous or polytomous logistic regression model in the health and life sciences. Multinomial Logistic Regression is the extension of the (binary) logistic regression (Chan, 2004) when the categorical dependent outcome has more than two levels. It gives a simultaneous representation (summary) of the odds being in one category relative to being in another category for all pairs of categories. Stepwise regression may be used in the exploratory phase of research but it is not recommended for theory testing (Menard, 1995). Theory testing is the testing of a-priori theories or hypotheses of the relationships between variables.

2.2.1 Chi-Square Automatic Interaction Detection (CHAID)

The most often- used criterion-based segmentation techniques are Automatic Interaction Detection (AID), CHAID and Classification and Regression Trees (CART) (Jonker, Franses & Piersma, 2002). The result of these three algorithms is a decision tree structure with a split at each node.

AID operates on an interval scaled dependent variable and maximizes the between-group-sum-of-squares (essentially the F-statistic) at each bisection. In contrast, CHAID operates on a nominal or ordinal scaled dependent variable and maximizes the significance of a chi-squared statistic at each partition, which need not be bisection (Kass, 1980). CART is preferred when there are many continuous variables and CHAID when there are many categorical variables.

CHAID is a technique that recursively partitions a population into separate and distinct sub-populations or segments such that the variation of the dependent variable is

minimised within the segments, and maximised among the segments. It works with all types of continuous and categorical variables. CHAID is concerned with predicting a single variable, the dependent variable, based on a number of other variables, referred to as predictor variables. It partitions the data into mutually exclusive, exhaustive, subsets that best describe the dependent variable (Kass, 1980). The subsets are constructed by using small groups of predictors. The dependent variable can have more than-two categories.

2.2.1.1 Basic tree-building algorithm

The Chi-squared test of independence is a non-parametric procedure, in that no distributional assumptions of the data need to be made (Diepen & Franses, 2006). Haughton & Oulabi (1997) studied the performance of response models built with CHAID; Bult & Wansbeek (1995) devised a profit maximisation approach to select customers by CHAID; and Levin & Zahavi (2001) studied CHAID using the logistic regression model as a benchmark for the comparative analysis.

CHAID method partitions a contingency table produced from cross-tabulation of three or more variables by using a semi hierarchical, sequential procedure. The procedure is semi hierarchical in the sense that it determines the smallest number of groupings (splits) of the levels of a predictor by a process of pair wise merging (and then separating) of the response levels on each of the predictors (Perreault & Barksdale, 1980).

The most significant relationships are used to control the structure of a tree diagram. Because the goal of classification trees is to predict or explain responses on a categorical dependent variable, the technique has much in common with the techniques used in the more traditional methods of discriminant analysis, cluster analysis, nonparametric statistics and nonlinear estimation.

Both CHAID and CART techniques will construct trees, where each (non-terminal) node identifies a split condition, to yield optimum prediction (of continuous dependent or

response variables) or classification (for categorical dependent or response variables). The final nodes (called leaves) are defined as combinations of the used independent variables or predictors.

The analyst should remember that CHAID is a multivariable procedure, but not a multivariate one. All variables are not considered simultaneously in the multivariable procedure.

Let the dependent variable have $d \geq 2$ categories, and a particular predictor under analysis $c \geq 2$ categories. A sub problem in the analysis is to reduce the $c \times d$ contingency table to the most significant $j \times d$ table by combining (in an allowable manner) categories of the predictor. Conceptually, we may first calculate statistics $T_j^{(i)}$, the usual χ^2 statistics for the ith method of forming a $j \times d$ table ($j = 2,3,\ldots,c$; the range of i depending on type of the predictor). Then, if $T_j^{(*)} = \max_i T_j^{(i)}$ is the χ^2 statistic for the best $j \times d$ table, choose the most significant $T_j^{(*)}$ (Kass, 1980).

Specifically, the full algorithm proceeds as follows:

Preparing predictors. The first step is to create categorical predictors out of any continuous predictors by dividing the respective continuous distributions into a number of categories with an approximately equal number of observations. For categorical predictors, the categories (classes) are "naturally" defined.

Merging Categories. The next step is to cycle through the predictors to determine for each predictor the pair of (predictor) categories that is least significantly different with respect to the dependent variable; for classification problems (where the dependent variable is categorical as well), it will compute a *Chi*-square test (Pearson *Chi*-square); for the regression problems (where the dependent variable is continuous), F tests. If the respective test for a given pair of predictor categories is not statistically significant as defined by an alpha-to-merge value, then it will merge the respective predictor categories and repeat this step (i.e., find the next pair of categories, which now may include

previously merged categories). If the statistical significance for the respective pair of predictor categories is significant (less than the respective alpha-to-merge value), then (optionally) it will compute a Bonferroni adjusted p-value for the set of categories for the respective predictor.

Selecting the split variable. The next step is to choose the split the predictor variable with the smallest adjusted p-value, i.e., the predictor variable that will yield the most significant split; if the smallest (Bonferroni) adjusted p-value for any predictor is greater than some alpha-to-split value, then no further splits will be performed, and the respective node is a terminal node.

This process is continued until no further splits can be performed (given the alpha-to-merge and alpha-to-split values). Perreault & Barksdale (1980) represent the CHAID algorithm in a flow chart (Figure. 2.1)

CHAID can be used to pre-screen data to exclude extraneous variables, that is, those with no predictive utility (Babinec, 1990). It reveals non-linearities and interactions in the explanatory variables.

2.2.1.2 Mathematical description of CHAID

CHAID (presented by Kass, 1975) is a natural measure since the nominal dependent variable and a categorized predictor allow the data to be summarized in a contingency table as in Table 2.1.

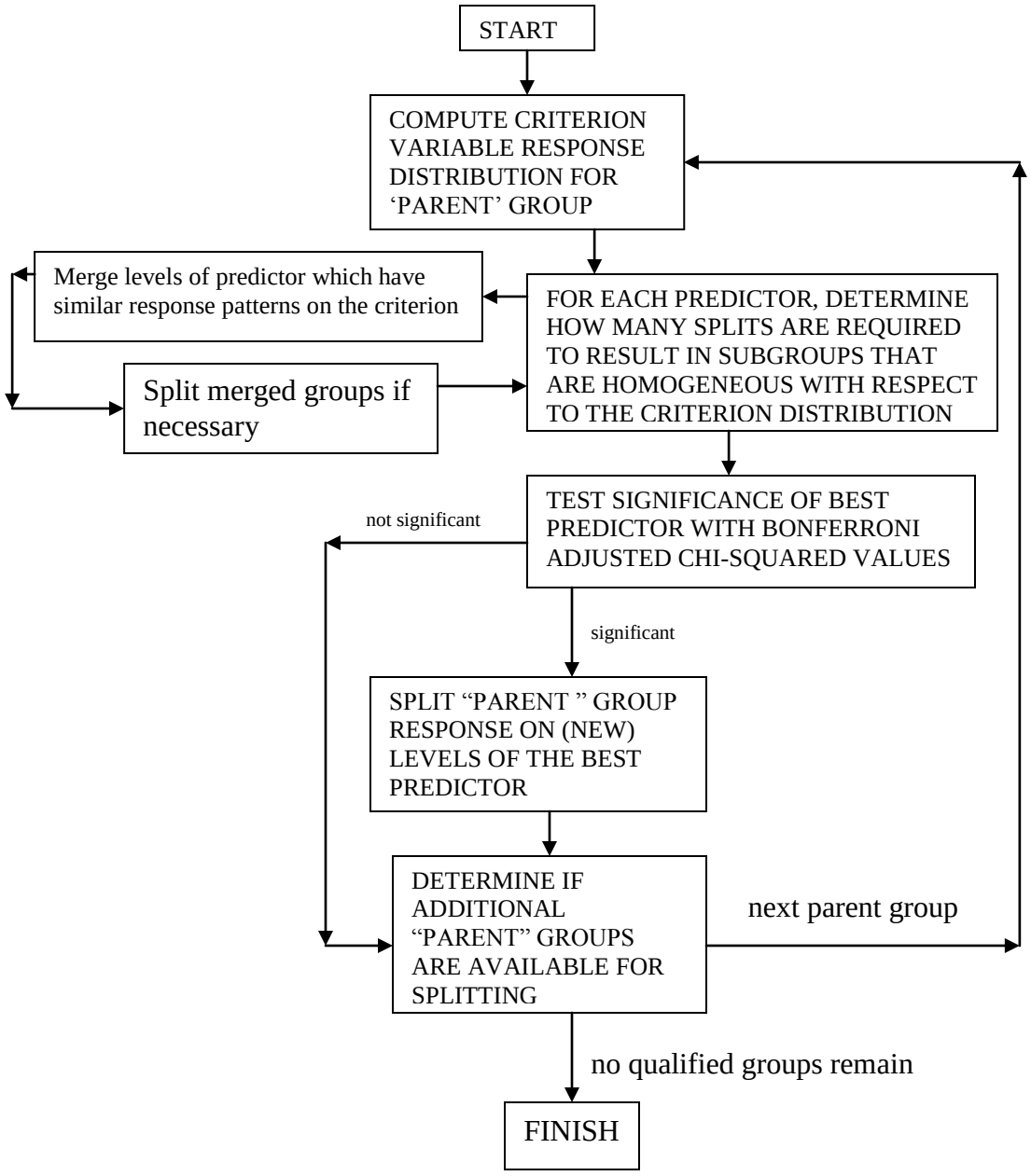

Figure 2.1: The CHAID algorithm Perreault & Barksdale (1980)

2.2.1.2.1 Outline of the technique

Table 2.1 Summary of data for one predictor (Kass, 1975).

		Dependent Variable					
		1	2	3	...	..	d
Predictor	1	O_{11}	O_{12}	O_{13}	...	..	O_{1d}
	2	O_{21}	O_{22}	O_{23}	...	..	O_{2d}
	3	O_{31}	O_{32}	O_{33}	...	..	O_{3d}
	.	..	..	..	..	..	..
	.	..	..	..	..	..	..
	c	O_{c1}	O_{c2}	O_{c3}	...	..	O_{cd}

O_{ij} denotes the number of observations that are classified into category i of the predictor and have characteristic j of the dependent variable.

Firstly, the "full" contingency table with c rows (for a c - category predictor) is calculated. Pairs of rows (as determined by the type of the predictor) are examined and the most insignificantly different pairs of rows are merged if their significance does not exceed some predetermined alpha That is, the various $2 \times d$ (for a d- category dependent variable) sub tables are examined and the least significant sub table "collapsed" unless its significance reaches some pre-assigned value.

The $(c\text{-}1) \times d$ table is similarly analyzed to determine if two further rows can be merged. When the merging criterion is not met, or when all rows have been merged, all compound rows are re-examined to determine if they can be split up again. That is, a compound row is examined to see if a $2 \times d$ table that can be formed from it, is significant at some pre-assigned value. This process is continued until no two rows are sufficiently similar to be merged, and no compound row can be decomposed into two significantly different rows. In order that such a stable solution should be reached it is necessary to ensure that the criterion for splitting a compound row is stricter than the

criterion for merging two rows. That is, the critical value for a split must have significance greater than that required to prevent a merge.

2.2.1.2.2 Case of a dichotomous dependent variable

Let the "reduced" 2×2 table be as in Table 2.2, where the first and second rows of the predictor are possible compound rows corresponding to the two parts of the split.

Table 2.2 Dichotomous dependent variable

		Dependent Variable		Total
		1	2	
Predictor	Row 1	a	b	a + b
	Row 2	c	d	c + d
	Total	a + c	b + d	N = a + b + c + d

The conventional chi-square statistic for this 2×2 table is (Conover, 1971)

$$T = \frac{N(ad - bc)^2}{(a+b)(c+d)(a+c)(b+d)} \tag{2.1}$$

Considering the dependent variable as a zero-one variable:-
The mean of the dependent variable:
$\bar{X} = (b+d)/N$

The variance of the dependent variable:

$$S^2 = \left((b+d) - N\overline{X}^2\right) \Big/ N$$

$$= \left((b+d) - N(b+d)^2 / N^2\right) \Big/ N$$

$$= (b+d)(1-(b+d)/N) \Big/ N$$

$$= (b+d)(N-(b+d))/N^2$$

$$= (b+d)(a+c) \Big/ N^2$$

Number of observations in the first group:

$M_1 = a + b$

Number of observations in the second group:

$M_2 = c + d$

The mean of the first group:

$$\overline{Y}_1 = b \Big/ (a+b)$$

The mean of the second group:

$$\bar{Y}_2 = d \big/ (c+d)$$

Hence from equation (2.1) the test statistic is

$$K = \left(\frac{M_1 M_2}{N}\right)\left(\frac{(\bar{Y}_1 - \bar{Y}_2)^2}{S^2}\right)$$

$$= \left(\frac{(a+b)(c+d)}{N}\right)\left(\frac{\left(\frac{d}{c+d} - \frac{b}{a+b}\right)^2}{(b+d)(a+c)\big/N^2}\right)$$

$$= \left(\frac{(a+b)(c+d)N^2}{N(b+d)(a+c)}\right)\left(\frac{(d(a+b) - b(c+d))^2}{((a+b)(c+d))^2}\right)$$

$$= \frac{N(ad-bc)^2}{(b+d)(a+c)(a+b)(c+d)} \qquad (2.2)$$

$$= T \quad \text{from (2.1)}$$

If the predictor is also dichotomous then T naturally has a chi-square distribution with one degree of freedom since it is derived from a 2×2 contingency table and K is likewise distributed in this case.

2.2.1.2.3 Some properties of the technique

Since the merging process compares two rows at a time this implies that, if they have the same proportional distribution on each category of the dependent variable, these two rows will be merged of necessity. Consider two such rows depicted in Table 2.3 in which $a_k/A = b_k/B = C_k$ say, for all k.

Table 2.3 A sub table for two rows

		Dependent variable						Total
		1	2	3	…	…	d	
Predictor	Row i	a_1	a_2	a_3	…	…	a_d	A
	Row j	b_1	b_2	b_3	…	…	b_d	B
Total		n_1	n_2	n_3	…	…	n_d	N

The chi-square for this sub table is calculated by

$$X = \sum_{k=1}^{d} \left(\frac{(a_k - n_k A/n)^2}{n_k A/n} + \frac{(b_k - n_k B/n)^2}{n_k B/n} \right). \qquad (2.3)$$

since the expected value in each cell is given by the product of the marginal totals divided by the grand total.

Since $a_k = AC_k$ and $b_k = BC_k$ it follows that

$n_k = a_k + b_k$

$= AC_k + BC_k$

$= (A+B)C_k$

$= nC_k$

So that

$a_k - n_k A/n = A\,C_k - n\,C_k A/n = 0$

and

$b_k - n_k B/n = B\,C_k - n\,C_k B/n = 0$

So that each numerator in the sum (2.3) is zero, and hence in this case $X = 0$, which is its minimum value and thus always not significant.

Since the test statistic is unaffected by a change in location of the dependent variable, take the middle two categories as having zero mean in order to simplify the algebra.

Let the sample sizes in each of the four groups be a, b, c and d respectively, each of these quantities being strictly positive. Let the mean of the first category be y_1, and of the last category y_2 where y_1 and y_2 are unconstrained. The four categories are depicted in Figure 2.2.

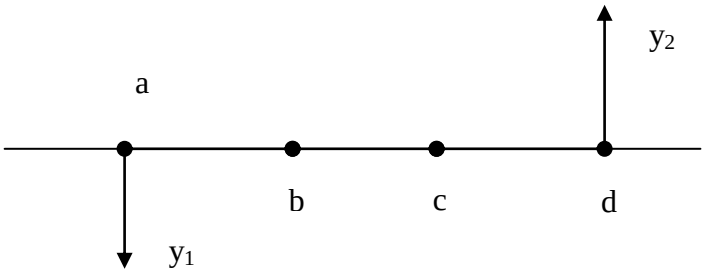

Figure 2.2: Four category example whose middle two categories have identical means.

The between-sum-of-squares is given by

$$B.S.S. = n_1\left(\overline{X}_1 - \overline{X}\right)^2 + n_2\left(\overline{X}_2 - \overline{X}\right)^2 \qquad (2.4)$$

where n_i and $\overline{X}_i$ are the number of observations and their mean in each part of the split and

$$\overline{\mathrm{X}} = \text{the grand mean} = \left(n_1\overline{X}_1 + n_2\overline{X}_2\right)/\left(n_1 + n_2\right) \qquad (2.5)$$

So

$$B.S.S. = n_1\bar{X}_1^{\,2} - 2n_1\bar{X}_1\bar{X} + n_1\bar{X}^2 + n_2\bar{X}_2^{\,2} - 2n_2\bar{X}_2\bar{X} + n_2\bar{X}^2$$

$$= n_1\bar{X}_1^{\,2} + n_2\bar{X}_2^{\,2} - 2\bar{X}\left(n_1\bar{X}_1 + n_2\bar{X}_2\right) + \left(n_1 + n_2\right)\bar{X}^2$$

$$= n_1\bar{X}_1^{\,2} + n_2\bar{X}_2^{\,2} - 2\bar{X}\left(n_1 + n_2\right)\bar{X} + \left(n_1 + n_2\right)\bar{X}^2$$

$$= n_1\bar{X}_1^{\,2} + n_2\bar{X}_2^{\,2} - \left(n_1 + n_2\right)\bar{X}^2 \qquad (2.6)$$

Note that the last term on the right of equation (2.6) is independent of how the group is split since $n_1 + n_2$ is the total sample size and $\bar{X}$ the grand mean, both of which remain unchanged.

If the split is after the first category, then in the above notation:

$$n_1 = a, \qquad n_2 = b + c + d, \qquad \bar{X}_1 = y_1, \qquad \bar{X}_2 = dy_2 \big/ \left(b + c + d\right),$$

and hence from (2.6)

$$(\text{B.S.S.})_1 = ay_1^{\,2} + d^2y_2^{\,2} / \left(b + c + d\right) - \left(n_1 + n_2\right)\bar{X}^2 \qquad (2.7a)$$

If the split is after the second category:

$$n_1 = a + b, \qquad n_2 = c + d, \qquad \bar{X}_1 = ay_1 \big/ \left(a + b\right), \qquad \bar{X}_2 = dy_2 \big/ \left(c + d\right),$$

and hence from (2.6)

$$\left(B.S.S.\right)_2 = a^2y_1^{\,2} / (a + b) + d^2y_2^{\,2} / (c + d) - (n_1 + n_2)\bar{X}^2 \qquad (2.7b)$$

Finally if the split is after the third category:

$n_1 = a + b + c,$ $\quad n_2 = d,$ $\quad \bar{X}_1 = ay_1/(a+b+c),$ $\quad \bar{X}_2 = y_2,$

and hence from (2.6)

$$(B.S.S.)_3 = a^2 y_1^{\,2}/(a+b+c) + dy_2^{\,2} - (n_1+n_2)\bar{X}^2 \qquad (2.7c)$$

If the split after the second category dominates both other splits, then

$(B.S.S.)_1 < (B.S.S.)_2$ and $(B.S.S.)_2 > (B.S.S.)_3$

That is

$$ay_1^{\,2} + d^2 y_2^{\,2}/(b+c+d) < a^2 y_1^{\,2}/(a+b) + d^2 y_2^{\,2}/(c+d) \qquad (2.8a)$$

and

$$a^2 y_1^{\,2}/(a+b) + d^2 y_2^{\,2}/(c+d) > a^2 y_1^{\,2}/(a+b+c) + dy_2^{\,2} \qquad (2.8b)$$

These equations give respectively

$$ay_1^{\,2}/(a+b) < d^2 y_2^{\,2}/((c+d)(b+c+d)) \qquad for \quad b \neq 0 \qquad (2.9a)$$

and

$$a^2 y_1^{\,2}/((a+b)(a+b+c)) > dy_2^{\,2}/(c+d) \qquad for \quad c \neq 0 \qquad (2.9b)$$

and hence

$$ay_1^2(c+d)/\left(dy_2^2(a+b)\right) < d/(b+c+d) \qquad (2.10a)$$

and

$$ay_1^2(c+d)/\left(dy_2^2(a+b)\right) > (a+b+c)/a \qquad (2.10b)$$

for $d>0$, $y_2 \neq 0$, $b>0$, $a>0$.

The equations (2.10a and 2.10b) together imply

$$(a + b + c)/a < d/(b + c + d)$$

which in turn implies

$$(a + b + c)(b + c + d) < ad$$

That is

$$a(b + c) + ad + (b + c)^2 + d(b + c) < ad$$

ie.

$$(b + c)(a + b + c + d) < 0 \qquad (2.11)$$

Clearly (2.11) is impossible since the sample size within each category is non-negative. Hence there must exist a better split than that separating two identical contiguous categories. A similar proof to the above shows that this result also holds in the three category case.

2.2.1.2.4 Convergence of the procedure

Referring to Figure 2.2, let the total chi-square for this table be denoted by T (with $(c-1)(d-1)$ degrees of freedom). Let the statistic for merging any two rows i and j (merging of course over all d columns) be denoted by X (with $d-1$ degrees of freedom). Given that these two rows i and j are merged let the resultant $(c-1) \times d$ table have chi-square Y (with $(c-2)(d-1)$ degrees of freedom). Hence the asymptotic result quoted above may be written as

$$T = X + Y \qquad (2.12)$$

Each time two rows are merged a new table results, and the process is repeated with the T of the new (reduced) table taking the value of the Y of the previous table.

For any merge, X is non-negative and hence by equation (2.12) the total chi-square for the succession of tables obtained by merging various pairs of rows is a non-increasing sequence $T_0, T_1, T_2, \ldots,$ say. Any sub table obtained from the original $c \times d$ table, must be a non-negative chi-square T_i that cannot exceed the original chi-square T_0, and hence is bounded. While the number of possible sub tables may be large, there is a finite number of them, and hence in any particular case T_i can only assume a finite number of values.

The procedure of merging and splitting produces a series of bounded values $\{T_i\}$ where $T_i \leq T_{i-1}$ if a merge occurs and $T_i \geq T_{i-1}$ if a split occurs.

A loop, if it exists, must contain the same number of merges as splits in order to return to an identical point in the analysis. Consider the series $\{T_i\}$ for such a loop. By assumption the T value will be identical at the beginning and end of the loop since the same sub table has been attained. However because of the merging criterion, each merge can reduce a particular T_i by at most some critical value k, say. On the other hand a split will increase a particular T_i by at least some critical value k', say, where the

algorithm demands $k' > k$. If there are m merges and hence also m splits in the "loop" the starting T value of the "loop" will be decreased by at most mk for the merges, and increased by at least $mk' > mk$ for the splits. Hence a contradiction results since the value for T at the end of the "loop" cannot coincide with the value at the beginning of the "loop". So it must be concluded that such "looping" is impossible.

For finite samples, the relationship $T = X + Y$ is only true approximately (except in some special cases to be considered in a moment) but Kendall & Stuart (1961, page 577) claim "the approximate partition is good enough for most practical purposes."

We briefly examine some details of the approximation (2.12) $T = X + Y$. Let the two rows that are merged be given by Table 2.3 so that

$$X = \sum_{k=1}^{d} \left(\frac{(a_k - An_k / n)^2}{An_k / n} + \frac{(b_k - Bn_k / n)^2}{Bn_k / n} \right) \qquad \text{from (2.3)}$$

$$= \frac{1}{nAB} \sum_{k=1}^{d} \frac{1}{n_k} \left(n^2 \left(a_k{}^2 B + b_k{}^2 A \right) - 2n_k nAB(a_k + b_k) + n_k{}^2 AB(A + B) \right)$$

$$= \frac{1}{nAB} \sum_{k=1}^{d} \frac{1}{n_k} \left(n^2 \left(a_k{}^2 B + b_k{}^2 A \right) - n_k{}^2 nAB \right)$$

$$= \frac{1}{AB} \sum_{k=1}^{d} \frac{1}{n_k} \left((A + B) \left(a_k{}^2 B + b_k{}^2 A \right) - (a_k + b_k)^2 AB \right)$$

$$= \frac{1}{AB} \sum_{k=1}^{d} \frac{1}{n_k} \left(a_k{}^2 AB + b_k{}^2 A^2 + a_k{}^2 B^2 + b_k{}^2 AB - a_k{}^2 AB - b_k{}^2 AB - 2a_k b_k AB \right)$$

$$= \frac{1}{AB} \sum_{k=1}^{d} \frac{1}{n_k} (a_k B - b_k A)^2 \qquad (2.13)$$

If the full table with chi-square T has column totals given by N_i (i = 1,2,…,d) and $N = \sum N_i$, then the sub table obtained from the total table by merging row i and row j is given by

$Y = T$-(contribution to chi-square from individual rows i & j) + (contribution to chi-square from the merged rows i & j)

That is

$$Y = T - \sum_{k=1}^{d} \left(\frac{(a_k - AN_k / N)^2}{AN_k / N} + \frac{(b_k - BN_k / N)^2}{BN_k / N} \right) + \sum_{k=1}^{d} \frac{(n_k - nN_k / N)^2}{nN_k / N}$$

So,

$$Z = \frac{1}{ABnN} \sum_{k=1}^{d} \frac{1}{N_k} \left[Bn(Na_k - AN_k)^2 + An(Nb_k - BN_k)^2 - AB(Nn_k - nN_k)^2 \right]$$

$$= \frac{1}{ABnN} \sum_{k=1}^{d} N_k^{-1} \begin{pmatrix} N^2 a_k^2 Bn + A^2 N_k^2 Bn - 2a_k NN_k ABn + N^2 b_k^2 An + B^2 N_k^2 An \\ - 2b_k NN_k ABn - N^2 (a_k + b_k)^2 - (A+B)^2 N_k^2 AB \\ + 2(a_k + b_k) NN_k nAB \end{pmatrix}$$

$$= \frac{1}{ABnN} \sum_{k=1}^{d} N_k^{-1} \begin{pmatrix} N^2 a_k^2 AB + N^2 a_k^2 B^2 + N_k^2 A^3 B + N_k^2 A^2 B^2 + N^2 b_k^2 A^2 \\ + N^2 b_k^2 AB + N_k^2 A^2 B^2 + N_k^2 AB^3 - N^2 a_k^2 AB - N^2 b_k^2 AB - \\ 2N^2 a_k b_k N_k AB - N_k^2 A^3 B - 2N_k^2 A^2 B^2 \end{pmatrix}$$

$$= \frac{1}{ABnN} \sum_{k=1}^{d} N_k^{-1} \left(N^2 a_k^2 B^2 + N^2 b_k^2 A^2 - 2N^2 a_k b_k AB \right)$$

$$= \frac{N}{ABn} \sum_{k=1}^{d} N_k^{-1} (a_k B - b_k A)^2 \qquad (2.14)$$

The error in the approximation $T = X + Y = X + (T - Z)$ is obtained from equations (2.13) and (2.14) as

$$Z - X = \frac{1}{AB}\sum_{k=1}^{d}\left(a_k B - b_k A\right)^2\left(\frac{N}{nN_k} - \frac{1}{n_k}\right)$$

$$= \frac{1}{AB}\sum_{k=1}^{d}\left(a_k B - b_k A\right)^2\left(\frac{n_k N - nN_k}{nN_k n_k}\right) \qquad (2.15)$$

Two special cases when this error is zero are immediately apparent. Firstly, if the two rows i and j each have the same proportion of observations within each category, chi-square X will be zero, which from (2.13) implies that $a_kB = b_kA$ for all k, and hence Z-X will be zero from (2.15). The second possibility is that $n_kN = nN_k$ for all k in (2.15) which implies that $n_k/N_k = n/N$, that is the column totals for the $2 \times d$ sub table and the column totals for the original $c \times d$ table are in constant proportion.

2.2.1.3 Validating tree results

Despite the lack of (objective) validation of tree outcomes and in particular CHAID outcomes in practice, there has been some attention for the topic in scientific literature. Arentze & Timmermans (2003) have developed a probabilistic classification method, where an individual in a specific segment is assigned to a certain class with a certain probability. Furthermore, they develop goodness-of-fit measures that give an indication of the likelihood of accurate prediction in a given sample of individuals. However, as the goodness-of-fit measures reflect the overall prediction quality of the tree, there is no indication how these measures could be used to assess the accuracy of the separate response percentages in each leaf-node of a CHAID tree (Diepen & Franses, 2006).

The bootstrap has also been applied to the decision tree structure before. Breiman (1996) developed bagging, an acronym for bootstrap aggregating.

2.2.1.4 Statistical distributions

CHAID is a non-parametric algorithm, which means that no distributional assumptions of the data have to be made. The only condition for CHAID to work effectively is that the data set used is large. Some indications on sample size can be found in the literature. For example Chaturvedi & Green (1995) mention a minimum of 1000-2000 cases. In CHAID, interactions in the data reveal themselves automatically. In a logit formulation, the researcher would have to subjectively choose which interactions to include.

2.2.2 Binary Logistic Regression

Logistic regression (Cox, 1970) is a type of predictive model that can be used when the target variable is a categorical variable. It does not involve decision trees and is more akin to nonlinear regression. Logistic regression can be used to predict a dependent variable on the basis of continuous and/or categorical independents and to determine the percent of variance in the dependent variable explained by the independent variables; to rank the relative importance of independent variables; to assess interaction effects; and to understand the impact of covariate control variables. Binary logistic regression is the two-group logistic regression model. Multinomial Logistic Regression is the extension for the (binary) logistic regression when the categorical dependent outcome has more than two response levels. It can be divided into two cases: ordinal response and nominal response.

For ordinal response, cumulative logits can be modeled with the proportional odds model. The proportional odds model assumes that the cumulative logits can be represented as parallel linear functions of independent variables, that is, for each cumulative logit the parameters of the models are the same, except for the intercept (Stokes, Davis & Koch, 2000)

If the proportional odds (parallel regression lines) assumption is not satisfied, the generalized logits approach can be used to model the relationship between the response

and independent variables. The generalized logit regression models are also used when the response variable is nominal (Stokes et al 2000).

For a categorical variable, the generalized logits are defined as natural logarithm, log, of the probability of each category over the probability of the last response category. These generalized logits are modeled as linear functions of independent variables with different regression parameters for each logit (not only intercepts as in the ordinal logistic regression, but all parameters are different) (Stokes et al 2000).

2.2.2.1 The model

Let *Y* be a binary dependent variable that assumes two outcomes or classes, typically labeled 1 with a probability of success P, or 0 with probability of failure *1- P*. The binary logistic regression model (BLR) classifies an individual into one of the classes based on the values for predictor (independent) variables $X_1, X_2, ...X_k$ for that individual. According to Hosmer & Lemeshow (2000) BLR estimates the *logit of Y*- a log of the odds of an individual belonging to class 1; the logit is defined in equation (2.16). The logit can be easily be converted into the probability of an individual belonging to class 1, *Prob(Y=1),* which is defined in equation (2.17).

$$\log it\, Y = b_0 + b_1 * X_1 + b_2 * X_2 + ... + b_k X_k \qquad (2.16)$$

$$\Pr ob(Y = 1) = \frac{\exp(\text{logit } Y)}{1 + \exp(\text{logit } Y)} \qquad (2.17)$$

The *b*s are the logistic regression coefficients.

Regression coefficients have a useful interpretation with a dummy dependent variable- they show the increase or decrease in the predicted probability of having a characteristic or experiencing an event due to a one-unit change in the independent variables. Instead of

least squares, logistic regression relies on maximum likelihood procedures to obtain the coefficient estimates.

For logistic regression, the procedure begins with an expression for the likelihood of observing the pattern of occurrences ($Y = 1$) and non-occurrence ($Y = 0$) of an event or characteristic in a given sample. It depends on unknown logistic regression parameters.

The maximum likelihood function in logistic regression is as follows (Pampel, 2000):

$$L = \Pi\left\{P_i^{Y_i} * \left(1 - P_i\right)^{1-Y_i}\right\},$$

where L refers to the likelihood, Y_i refers to the observed value of the dichotomous dependent variable for case i, and θ_i refers to the predicted probability for case i. The logit equals

$$L_i = \ln\left[P_i / \left(1 - P_i\right)\right]$$

The key is to identify b values that produce L_i and θ_i values that maximize L. Taking the natural log of both sides of the likelihood equation gives the log likelihood function:

$$\ln L = \sum \left\{\left[Y_i * \ln P_i\right] + \left[\left(1 - Y_i\right) * \ln\left(1 - P_i\right)\right]\right\}.$$

There will be $k + 1$ likelihood equations that are obtained by differentiating the log likelihood function with respect to the $k + 1$ coefficients. The likelihood equations that result may be expressed as follows:

$$\sum_{i=1}^{k}\left[y_i - P_i\right] = 0$$

and

$$\sum_{i=1}^{k} x_{ij}\left[y_i - P_i\right] = 0$$

for $j = 1, 2, \ldots, k$.

2.2.2.2 Tests of Significance

The logistic regression in SAS calculate the Wald statistic for a (two-tailed) test of a single coefficient, which equals the square of the ratio of the coefficient divided by its standard error and has a chi-square distribution. Raftery (1995) has recently proposed a Bayesian information criterion (BIC) for a variety of statistical tests. Specifically, the BIC value refers to the difference in model information with and without the variable and coefficient in question. If the BIC value for a variable equals or falls below 0, the data provide little support for including the variable in the model. He defines a BIC difference of 0-2 as weak, 2-6 as positive, 6-10 as strong, and greater than 10 as very strong.

The base line log likelihood comes from including only a constant term in the model-the equivalent of using the mean probability as the predicted value for all cases. Multiplying the difference between the baseline log likelihood and the model log likelihood by -2 gives a chi-square value with degrees of freedom equal to the number of independent variables (not including the constant, but including squared and interaction terms). For a given degree of freedom, the larger the chi-square value, the greater the model improvement over the baseline, and the less likely that all the variables coefficients equal 0 in the population.

2.2.3 Multinomial Logistic Regression

When one considers a regression model for a discrete outcome variable with more than two responses, one must pay attention to the measurement scale. In this section we discuss the logistic regression model for the case in which the outcome is a nominal

scale. It is another method for performing classification. For nominal response outcomes, we form generalized logits and perform a logistic analysis. The basic concept was generalized from binary logistic regression (Aldrich & Nelson 1984, Hosmer & Lemeshow 2000). In a Multinomial Logistic Regression model, the estimates for the parameters can be identified compared to a baseline category (Long, 1997). The Multinomial Logistic Regression model with a baseline category would be expressed as follows:

$$Log\left(P_i \big/ P_1\right) = \alpha_i + \beta_i x, \quad i = 1,.....,I-1.$$

The logistic model uses the baseline-category logits with a predictor x. This model provides several equations for classifying individuals into one of many groups. The number of equations is one less than the number of groups. Each equation looks like the binary logistic regression model.

2.2.3.1 Fitting the Multinomial Logistic Regression model

When the class dependent variable takes on more than two outcomes or classes, the multinomial regression model, an extension of BLR model, can be used to predict class membership. We consider Y with 3 categories, coded 1, 2 and 3. The obvious extension is to use $Y = 1$ as the referent or baseline outcome and to form logits comparing $Y = 2$ and $Y = 3$ to it.

To develop the model, assume we have k covariates and a constant term, denoted by the vector, x, of length $k+1$ where $x_0 = 1$.

We denote the two logit functions as (Hosmer & Lemeshow, 2000)

$$g_2(x) = \ln\left[\frac{P(Y=2\mid x)}{P(Y=1\mid x)}\right]$$
$$= b_{10} + b_{11}x_1 + b_{12}x_2 + \ldots + b_{1k}x_k$$
$$= x'b_1 \qquad (2.18)$$

and

$$g_3(x) = \ln\left[\frac{P(Y=3\mid x)}{P(Y=1\mid x)}\right]$$
$$= b_{20} + b_{21}x_1 + b_{22}x_2 + \ldots + b_{2k}x_k$$
$$= x'b_2 \qquad (2.19)$$

It follows that the conditional probabilities of each outcome category given the covariate vector are

$$P(Y=1\mid x) = \frac{1}{1+e^{g_2(x)}+e^{g_3(x)}}, \qquad (2.20)$$

$$P(Y=2\mid x) = \frac{e^{g_2(x)}}{1+e^{g_2(x)}+e^{g_3(x)}}, \qquad (2.21)$$

and

$$P(Y=3\mid x) = \frac{e^{g_3(x)}}{1+e^{g_2(x)}+e^{g_3(x)}}, \qquad (2.22)$$

Each probability is a function of the vector of *2(k+1)* parameters $b' = \left(b_1', b_2'\right)$. A general expression for the conditional probability in the three category model is

$$P(Y=j\mid x) = \frac{e^{g_j(x)}}{\sum_{p=1}^{3} e^{g_p(x)}}$$

where the vector $b_0 = 0$ and $g_1(x) = 0$.

There are *n-1* logits for a MLR with *n* classes.

To construct the likelihood function we create three binary variables coded 0 or 1 to indicate the group membership of an observation. We note that these variables are introduced only to clarify the likelihood function and are not used in the actual multinomial logistic regression analysis. The variables are coded as follows: if *Y = 1* then $Y_1 = 1$, $Y_2 = 0$, and $Y_3 = 0$; if *Y = 3* then $Y_1 = 0$, $Y_2 = 1$, and $Y_3 = 0$; and if *Y = 2* then $Y_1 = 0$, $Y_2 = 0$, and $Y_3 = 1$. We note that no matter what value *Y* takes on, the sum of these variables is $\sum_{j=1}^{3} Y_j = 1$. Using this notation it follows that the conditional likelihood function for a sample of *n* independent observations is

$$L(b) = \prod_{i=1}^{n} \left[P_1^{\,y_{1i}} P_2^{\,y_{2i}} P_3^{\,y_{3i}} \right].$$

Taking the log and using the fact that $\sum y_{ij} = 1$ for each *i*, the log likelihood function is

$$l(b) = \sum_{i=1}^{n} y_{1i} g_1(x_i) + y_{2i} g_2(x_i) - \ln\left(1 + e^{g_1(x_i)} + e^{g_2(x_i)}\right) \qquad (2.23)$$

The likelihood equations are found by taking the first partial derivatives of *L(β)* with respect to each of the *2(k+1)* unknown parameters. The general form of these equations is:

$$\frac{\partial L(b)}{\partial b_{jk}} = \sum_{i=1}^{n} x_{pi}\left(y_{ji} - P_{ji}\right) \qquad (2.24)$$

for *j = 2,3* and *p = 0,1,2,...,k*, with $x_{0i} = 1$ for each subject.

The maximum likelihood estimator, $\hat{b}$, is obtained by setting these equations equal to zero and solving for b.

The matrix of second partial derivatives is required to obtain the information matrix and the estimator of the covariance matrix of the maximum likelihood estimator. The general form of the elements in the matrix of second partial derivatives is as follows:

$$\frac{\partial^2 L(b)}{\partial \beta_{jp} \partial b_{jp'}} = -\sum_{i=1}^{n} x_{p'i} x_{pi} P_{ji} \left(1 - P_{ji}\right) \tag{2.25}$$

and

$$\frac{\partial^2 L(b)}{\partial \beta_{jp} \partial b_{j'p'}} = -\sum_{i=1}^{n} x_{p''i} x_{pi} P_{ji} P_{j'i} \tag{2.26}$$

For j and $j' = 2,3$ and p and $p' = 0,1,2,...,k$. The observed information matrix, $I\left(\hat{b}\right)$, is the *2(k+1)* by *2(k+1)* matrix whose elements are the negatives of the values in equations (2.25) and (2.26) evaluated at $\hat{b}$. The estimator of the covariance matrix of the maximum likelihood estimator is the inverse of the observed information matrix,

$$\hat{Var}\left(\hat{b}\right) = I\left(\hat{b}\right)^{-1}.$$

Let the matrix X be the n by $k+1$ matrix containing the values of the covariates for each subject, let the matrix V_j be the n by n diagonal matrix with general element $\hat{P}_{ji}\left(1 - \hat{P}_{ji}\right)$ for $j = 2,3$ and $i = 1,2,3,...,n$, and let V_3 be the n by n diagonal matrix with general element $\hat{P}_{1i}\hat{P}_{2i}$. The estimator of the information matrix may be expressed as

$$\hat{I}\left(\hat{b}\right)=\begin{bmatrix}\hat{I}\left(\hat{b}\right)_{11} & \hat{I}\left(\hat{b}\right)_{12}\\ \hat{I}\left(\hat{b}\right)_{21} & \hat{I}\left(\hat{b}\right)_{22}\end{bmatrix}, \tag{2.27}$$

where

$$\hat{I}\left(\hat{b}\right)_{11}=\left(X'V_1X\right),$$

$$\hat{I}\left(\hat{b}\right)_{22}=\left(X'V_2X\right),$$

and

$$\hat{I}\left(\hat{b}\right)_{12}=\hat{I}\left(\hat{b}\right)_{21}=-\left(X'V_3X\right).$$

2.2.3.2 Interpreting the fit and odds ratio

We assume that the outcome labeled with *Y=1* is the reference outcome. The subscript on the odds ratio indicates which outcome is being compared to the reference outcome. The odds ratio of outcome $Y = j$ versus outcome $Y = 1$ for covariate values $x = a$ versus $x = b$ is

$$OR_j\left(a,b\right)=\frac{P\left(Y=j\mid x=a\right)/P\left(Y=1\mid x=a\right)}{P\left(Y=j\mid x=b\right)/P\left(Y=1\mid x=b\right)}.$$

When the covariate is binary, coded 0 or 1, we simplify the notation further and let $OR_j = OR_j(1,0)$.

Continuous covariates that are modeled as linear in the logit have a single estimated coefficient in each logit function. This coefficient, when exponentiated, gives the estimated odds ratio for a change of one unit in the variable.

2.3 Student Performance

Students entering South African universities come from a wide range of social and cultural backgrounds that give them very different life experiences, different educational opportunities, and a great variety of expectations, needs and academic potential (Chikte & Brand, 1996; Goduka, 1996). Despite the changing characteristics of those aspiring to attend University, the general entry requirements for undergraduate programmes in South Africa Universities have changed little in the past ten years. The relationship between school results, first-year University performance and the results of other assessment instruments has been of international interest for many years. It seems that only a very small percentage of those students with a school result of below 70% obtain a first-year University average performance of 50% or more (Roux, Bothma & Botha, 2004). Power, Robertson, & Baker (1987) stated "the stress should not only be on admitting a wider range of students, but also on giving them the support and help needed to ensure a reasonable chance for success" (Page 3). Study skills have also been found to influence academic performance.

One of the more vexing problems within higher education has been the unacceptably high number of first-year students who do not persist into the second year. However, the vast majority of students attend institutions with the highest attrition rates. The theories formulated by both Spady (1970) and Tinto (1975) identify several levels of independent variables that affect student retention, with only a few major variables having a direct effect on retention and the remaining variables affecting retention indirectly.

Student entry characteristics affect the level of initial commitment to the University. These student characteristics include family background characteristics (e.g., socio economic status, parental educational level), individual attributes (e.g., academic ability, race and gender) and previous institution experiences (e.g., high-school academic achievement) (Braxton, Milem & Sullivan, 2000).

Sex of the individual also appears to be related to University persistence with a higher proportion of men finishing University degree programs than women (Astin, 1972; Cope, 1971; Spady, 1970), but of those who drop out, a greater proportion of women tend to be voluntary withdrawals rather than academic dismissals (Lembesis, 1965; Robinson, 1969; Spady, 1971). Women tended to get better grades than the men during the first year. A man is about twice as likely as a woman to obtain borderline or failing grades as a University first year student. However, Astin (1997) has found that sex only explains 2% of the variance in retention.

Both Ishler and Upcraft (2005) and Stage and Hossler (2000) point out that racial/ethnic identity is a very difficult variable to cleanly assess due to the confounding interactions that occur between it and many other variables.

It is clear that performance in high school has been shown to be an important predictor of future University performance (Astin, 1971). Moreover, since it is also clear that the characteristics of the high school, such as its facilities and academic staff, are important factors in the individual's achievement (Dyer, 1968); it follows that they would also affect the individual's performance and therefore persistence in University.

The matriculation mark is a reasonably good predictor of pass/fail at University (Mitchell et al, 1997). The matriculation examination, initially set by the Joint Matriculation Board (JMB) and later by the four provinces of the previous governmental dispensation under the jurisdiction of the JMB, 'soon established itself as the only school-leaving certificate and gateway to the Universities and to many professional careers, and also was recognized by several foreign bodies' (Lolwana, 2004). Matric examination performance has been shown to be a reliable predictor of University performance for students from all education departments who are admitted on the basis of the results (Stoker, 1985; Jawitz 1995). For this reason matric examination performance has traditionally been used as a basis for the University entry requirements. A number of South African studies have pointed to the fact that on entry to higher education institutions, large numbers of

students are not sufficiently ready (i.e., do not have the required, academic, cognitive and personal competencies) to cope with higher education studies (Yeld, 2003).

In the analysis of 2004 first-year students in the various faculties at Nelson Mandela Metropolitan University (NMMU), the size of the correlation coefficients varies across Faculties (Foxcroft & Stumpf, 2005). The highest correlation was found for Pharmacy students. The entrance requirements for the Pharmacy programme are more stringent than for any other programme at the NMMU in that a high level of Matric performance (65%) and Mathematics performance is required for entry. Although the entry criteria for degree programmes in Science and Business and Economic Sciences are essentially the same, the correlation with academic performance for Science programmes is far worse than that for Business and Economic Science programmes.

There has been a strong call for Matric examination papers to be available in more of South Africa's eleven official languages. The impact of such mother-tongue education may be on the performance of students in higher education where the medium of instruction in the majority of institutions is English.

According to Amoore (2001), matriculation simply means 'University admission'. Currently, if a candidate's curriculum (subject groupings and Higher Grade/ Standard Grade (HG/SG) requirements) and results are in accordance with prescribed regulations, the candidate obtains a Senior Certificate with matriculation endorsement. Such an endorsement is the minimum statutory requirement for entry into bachelor's degree studies (Foxcroft and Stumpf, 2005).

Perhaps the most common educational justification for ability tracking is the assumption that the student will develop better academically if he is grouped with students of similar ability. The teaching practices of University hold importance for both our understanding of the process of University student departure and for the improvement of institutional retention rates (Braxton, et al, 2000).

Stage and Hossler (2000) suggest that parental educational attainment and yearly income are related to retention. After reviewing the the literature on student persistence, Pascarella and Terenzini (2005) conclude that the grades earned during the first year of college “may well be the single best predictors of student persistence”, even after taking into account students’ entering characteristics.

CHAPTER 3

METHODOLOGY

According to the pass records from 1995 to 2006, in the Faculty of Commerce at the University of the Witwatersrand, approximately 47% of first year students pass into their second year of study (Oracle Student System). On average 19% of first year students have been required to repeat the first year and 34% have been excluded. The purpose of this research is to find some of the important predictor variables of first year students' performance in the Faculty of Commerce, University of the Witwatersrand.

In this study theoretical models that can take into account a variety of attributes and pre-University experience that may affect a student's performance in the first year are suggested. To evaluate these models and explore the possible relationships between some of these variables and student performance, data were extracted from the student data base on the Oracle Student System at the University of the Witwatersrand. The data were cleaned for the purpose of this analysis. If the errors were not corrected, the analysis could come to wrong conclusions. CHAID and multinomial logistic regression analysis models were built using SAS software. The results of those two analyses are compared.

This chapter will detail the methods used in this research including the data source, data cleaning and the methods of data analysis.

3.1 Data Source

Data were collected on full-time, Faculty of Commerce, University of the Witwatersrand first year students from the four cohorts of 2003 to 2006. The University has a total enrolment of over 5,000 full time first year students per year in all undergraduate degrees in all faculties.

Data for the study were obtained from records maintained by the Management Information Unit (MIU) in the Oracle Student System (OSS). It provides management and statistics for the University of the Witwatersrand. OSS has been the system from 2 January 2007 at the University. This data system replaced the Student Information Record System (SIRS). The SIRS system had been the management information system of the University for more than twenty years. The migration from SIRS to OSS, involved some eight million records and 370,000 programme attempts. The migrated data had been validated. Some of the academic programmes have many course (unit) prerequisites and co-requisites. SIRS did not validate for these requirements.

The study was focussed only on the Faculty of Commerce and hence this was the only data extracted from the system. The student number is the unique identification of the student at the University. It was used as the identification variable to extract the students' information, and as the number is unique to each student this avoids repetition.

OSS contains the details of the students' personal identity number, student number, title, first name, last name, date of birth, email address, mobile number, home residence telephone number, marital status, religion, home language, postal address, gender, race, nationality, permit type, programme by organization, qualification type, calendar instance year, year of study, matriculation subjects and grades, previous institution type and performance at University. There is little value in analysing some of these variables, for example, personal identity number, address, phone number etc. Race was also not included as the students' background and matriculation records only were considered in this research. Because there are very few foreign students in the Faculty of Commerce,

nationality was also not considered. Further as the study was dealing with students who had in general just finished their secondary schooling and were usually 18 or 19, marital status was also not considered.

Thus the extracted database included the student number, programme code, students' first year overall University performance, birth date, gender and matriculation course grades for selected subjects and marks.

3.2 Data Cleaning

Data were initially separated by year (cohort). Part time and occasional students were excluded. If the students come from other previous post school institutions, they were possibly replicated with the same details two or more times. These duplicate records were deleted after every student's details were checked on a student by student basis.

The students enrolled in the Bachelor of Economic Science programme were also deleted from the analysis as their entry requirements and curricula are materially different from the other undergraduate programmes in the Faculty of Commerce. This analysis was only for the B. Com. (Ordinary) degree and B.Acc. (Ordinary) degree.

Unnecessary entries for the analysis have been deleted from the database. Every dataset contains some errors. Errors and inconsistencies have been first detected, identified and corrected where possible. Some students' details didn't give usable information. For example their matriculation subjects' details were missing. Those students and their records were also deleted from the data base. There was no imputation for the missing values and the data set, before the deletion of records contained less than 1% of missing values.

Two other variables were created: matriculation aggregate and age.

The matriculation aggregate for students with matriculation subjects only at the higher grade was calculated by averaging the marks received for the subjects passed. The aggregate for students who matriculated with some subjects at the standard grade level required a different calculation. A mark of 20 less than the standard grade mark was substituted for each mark obtained at a standard grade level. This is consistent with the admission policy of the Faculty of Commerce where Standard Grade 'A' has been treated as equivalent to a higher Grade "C" (admission requirements, www.wits.ac.za). Then the aggregate of all these results was taken. If a student registered for more than six subjects, only the six highest marks were taken into consideration provided that Mathematics and English were among these six subjects. After this the aggregate was then calculated. Only the common subjects Accountancy, Biology, English, History, Mathematics and Physical Science were included in the data base, and the remainder were then deleted. It is of little purpose in this analysis to have too few students with unusual subjects that the majority of students have not taken.

The student's age on registration is calculated from the student's date of birth and the year of registration.

Every student's records were checked again; the high school records of a small number of students were also missing from the data base (importantly English and Mathematics). Those students' records were also removed from the database and the analysis as they did not have information that was essential for the modelling.

In order to preserve student anonymity the student number was replaced. It was replaced by the natural number in order to protect the identity of the individual student.

Finally 4,787 records over the four cohorts had been created for the analysis. The data were in Excel format, and then were converted to SAS.

The variables used in the analysis and a code sheet for the variables to be considered in the CHAID and the logistic regression are given in Table 3.1.

Table 3.1 Variables and associated response categories

Variable	Codes
Sex	Male =0 Female=1
Age	Years
Previous institution type	FE(Further Education)=1 HE(Higher Education)=2 School=3
Accounting	HG=1 Not taking=2 SG=3
Biology	HG=1 Not taking=2 SG=3
English	First language=1 second language=2
History	HG=1 Not taking=2 SG=3
Mathematics	HG=1 SG=2
Physical Science	HG=1 Not taking=2 SG=3
Aggregate	Continuous variable
Student Performance	Completed=1 Returned=2 Excluded=3

Completed means that the students pass into their second year; returned means that students are enrolled in the Faculty but have not achieved sufficient marks to pass to the

second year of study i.e. they are required to repeat first year; excluded means that students who have cancelled their registration, or have been excluded either for financial or academic reasons. This category includes drop-outs.

Astin (1997) has concluded that these variables (Students' high school grades and gender) account for much of the variance in retention that can be predicted from entering first year student characteristics.

Gender was included in this study even though Astin (1997) had found that this variable explains less than 2% of the variance in retention.

Students' performance is chosen as the response variable with three categories, completed (pass into their second year), returned (still busy in the first year) and excluded (have cancelled the registration from the Faculty) and whereas the other variables- age, gender, previous institution type, grade of matriculation courses and school aggregate- are selected as predictor variables. The six school subjects of analysis for this study are English, History, Mathematics, Accountancy, Biology and Physical Science. All students take English at higher grade as first language or second language and Mathematics as higher grade or standard grade and these two school subjects with certain minima are the minimum requirements for the admission to the Commerce Faculty. History, Biology, Accountancy and Physical Science are the most popular courses that potential Commerce Faculty students take at school.

Age and matriculation Aggregate are continuous variables and the others are nominal variables. Only 56.7% of students were passed into their second year in the entire sample. The students that comprised the data set had an average age of 19.4 years and standard deviation 1.31 years upon enrolment, with a range of 15 to 36 years; and had an average matriculation aggregate of 63.9%, with a standard deviation of 9.95% and the average was between 41% and 96%.

3.3 Data Analysis

To carry out this research, CHAID and Multinomial Logistic Regression were used to analyse the students' performance data of the Faculty of Commerce, University of the Witwatersrand. Once a model has been built, its predictive value needs to be assessed. It is a most important step in the model building sequence. The data was partitioned into two data sets: A training data set (2003 & 2004 cohorts) and a validation data set (2005 & 2006 cohorts). The training data was used to estimate the model and the validation data was used to estimate the fitted model by assessing the misclassification rate. In the whole data set 48.1% of the students were females, whereas 47.4% of the training data were females.

This study used both categorical and continuous independent variables. The analysis of the data was conducted in two parts. The first part involved the fitting of models using CHAID and Multinomial Logistic Regression on the training data. In the second part of the study, models from CHAID and Multinomial Logistic Regression were validated using the validation datasets. The software package SAS was used to analyse the data.

CHAID was carried out using SAS Macros (CHAID Macro).The training data was in the Excel format. So, SAS training data set was created to use in the SAS. CHAID.sas macro-call file was opened from the mac-call folder in the Data mining using SAS applications CD-ROM into the SAS program editor window. The macro-call file CHAID.sas was submitted by clicking the run icon to open the macro-call window called CHAID. It contains ten parameters such as input the SAS data set name, response variable name, nominal variables etc., for creating CHAID decision tree diagrams and classification plots in that window. The appropriate parameters SAS data set name, response variable, nominal variables, ordinal variables, path of Xmacro.sas and display were input and run to submit the macro. The output gave the decision rule and decision tree.

Here, a decision tree partitions data into smaller segments called terminal nodes or leaves that are homogeneous with respect to a target variable. This partitioning goes on until the subsets cannot be partitioned any further using one of many user-defined stopping criterion. For the purpose of the CHAID analysis, only variables having a Bonferroni adjusted *p* value of less than or equal to 0.1 are eligible for segmentation. Any cell that had less than 5 subjects is ineligible for further segmentation. The classification summary for the training data was displayed in a donut chart (Fernandez, 2003).

The decision tree was cleaned as it was too large and complex (looping and with very small frequency cells), i.e. the decision tree was generated manually using the decision rule generated by the CHAID macro. The illustration of the decision tree was generated by smart draw program. The classification tree shows which variables may be used in further analysis and what variables may possibly be discarded.

The multinomial logistic model was calibrated with SAS statistical software. Here, for the students' performance, generalised logits were formed and the analysis was performed. The analysis of generalized logits is a form of the loglinear model. CATMOD procedure in SAS was used to model the generalised logits. A logit is formed for the probability of each succeeding students' performance over the last students' performance (returned). First, the SAS training data set was created and the CATMOD procedure was invoked. Maximum likelihood estimation was used to estimate the parameters.

There are a lot of variables and interactions in the training and validation data sets. It was too large for it to be put together in the SAS program. Initially the program was run for the single predictors and then two way interactions separately as it could not run all the three way interactions concurrently. Significant predictors and two way interactions were found. As a further step all the possible three way interactions were found and they were entered into the program. This was continued until all possible significant interactions were found. Finally, the significant predictors and interactions were put together to fit the model. After fitting the specified model, goodness-of-fit statistics was examined and

model reduction was performed. Sets of parameters (intercept and regression) were estimated for two logit functions completed versus returned and excluded versus returned. Observed and predicted frequencies were calculated using the fitted model.

After CHAID and Multinomial Logistic Regression models had been fitted to the training data set, they were validated using the independent validation data set. Validation means that it validates or confirms the derived models obtained from the training data.

The CHAID model was validated by applying the classification criteria to the independent validation data set and verifying the success of classification. First, SAS validation data was created. CHAID macro window was called as mentioned above. Options (valid SAS data set name) are available for validating the decision tree using independent validation data sets in the CHAID macro. The output presented misclassification error rates and measured the success of classification using validation data. The classification summary for the validation data was also displayed in a donut chart. An overall error rate was computed by an average error rate weighted by the students' performance frequency.

The validation data were classified manually according to the fitted classification model and tree was drawn. Completed percentages of every node compared with the fitted classification tree.

Multinomial Logistic Regression was validated using goodness-of-statistics. The likelihood ratio test was used as a goodness-of-fit test in this analysis. High p-values indicate adequate fit. Validation SAS data set was created. Fitted Multinomial Logistic Regression model using training data also was applied on validation data set to validate the model. Observed and predicted frequencies were calculated and overall error rate was determined.

CHAPTER 4

RESULTS

This research used data from the records of the Faculty of Commerce, University of the Witwatersrand to examine the academic performance of first year students. The statistical techniques, CHAID and Multinomial Logistic Regression were used to analyse the data. This chapter presents the results of these two analyses. The first section presents the results of the CHAID analysis and the next section is the Multinomial Logistic Regression.

4.1 CHAID

To facilitate the analysis, the continuous predictor variables are categorised into meaningful intervals based on the content domain of the problem under study. In this case, the intervals do not divide at the value which would lead to a perfect classification. However, continuous predictor variables are automatically binned into ordinal classes for the purpose of the analysis by the CHAID macro in SAS. Only adjacent categories of ordinal or grouped continuous predictors are allowed to merge.

The model was built using the training data and it was validated using the validation data set.

The purpose of the study is to examine the academic performance of first year students of the Faculty of Commerce, University of the Witwatersrand. Figure 4.1 gives the CHAID tree for the students' performance. The most salient dimension for students' performance in this data set is the matriculation Aggregate. The test of independence between aggregate and the students' performance yields the lowest error probability (0.0001). It explains more of the variation in students' performance than any other predictors included in the analysis. In Figure 4.1 the coding scheme used is: "c" represents the completed (pass into their second year), "r" represents the returned (still busy with the first year) and "e" represents the excluded (have cancelled or have been cancelled for financial, personal or academic reasons). In addition, the terminal nodes are numbered consecutively immediately below the cell.

At the root node a completed rate is 51.1% that is 51.1% of first year students have passed into their second year. It differs from the average completed rate of the cohorts from 1995 to 2006 because of the intra cohort variability. The number of student enrolment also differs from year to year. Furthermore separate marks of aggregate are merged, as there are no significant differences between them when related to the values of the criterion variable: students' performance. The sample is not divided at the child node with the value of the aggregate, 68-73. The nodes of the classification tree grouped by aggregate 41-56, 57-67, 68-73 and 74-94 are separated by the variables: Previous Institution Type, Accountancy, Physical Science, English and Age. This indicates that the remaining variables not listed in the classification tree have no significant influence to the prediction of students' performance in the model when using CHAID. It has to be considered, that Aggregate is a continuous variable and therefore only adjacent categories can be merged. In this analysis six out of ten variables are significant.

It should be noted that the completed percentage trends upward across the groups: Students with aggregates between 41-56 with a total sample of 720 is at a pass rate of 35.6%; 1039 students with aggregates between 57-67 is at a pass rate of 43.7%; students with aggregate 68-73 with the total sample size of 353 is at a pass rate of 64.3%; while 401 students with aggregates 74-94 is at a pass rate of 86.8%.

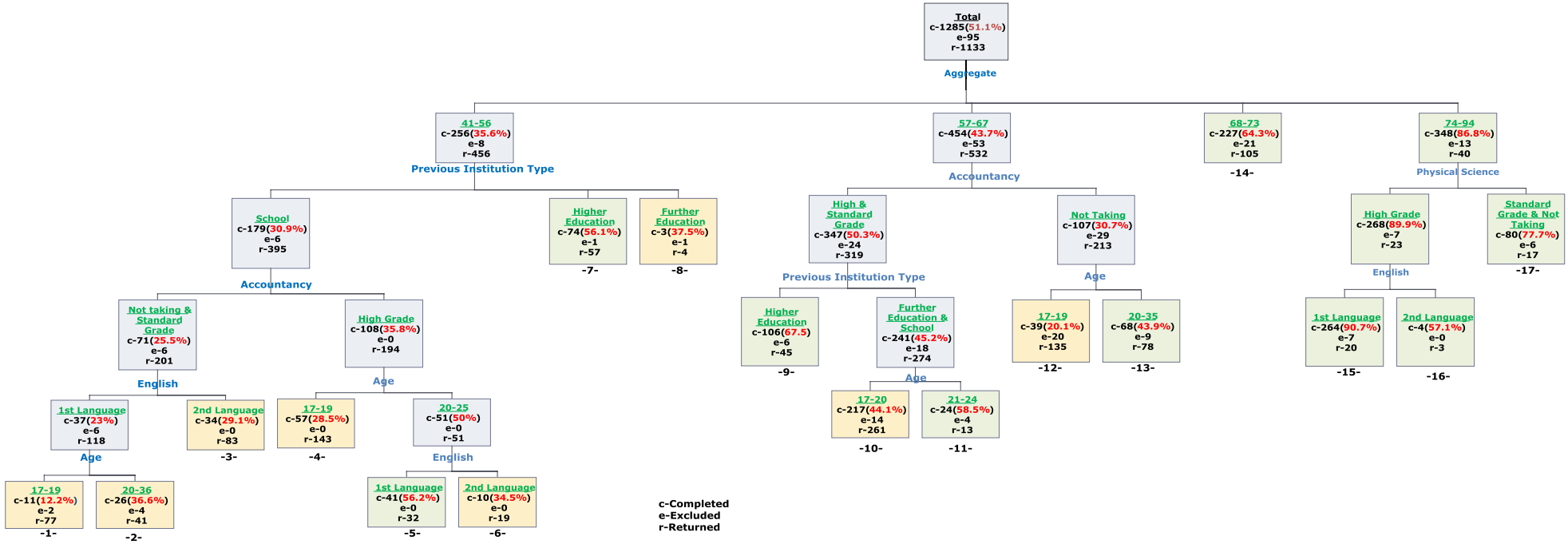

Figure 4.1 CHAID output for students' performance on the training data

The next split in the classification tree reveals is an interaction effect. At the second level of partitioning it was found that Previous Institution Type, Accountancy and Physical Science were the most important predictors.

At the third level of partitioning Accountancy, Previous Institution Type, Age and English all contributed to the variation in Students' performance. The fact that each of these variables explain most of the variation in respect of a specific subpopulation, points to the presence of interactions between these predictors and the preceding predictors. It is shown at every branch.

The completed pass rate of the students who have an aggregate between 41-56 is 35.6%. The next meaningful variable is previous institution type. It should be noted that students from school have 30.9% chance to pass into their second year and this sample size is 580; the pass rate of those who come from higher education is 56.1% and this sample size is 132; whereas the students from further education have 37.5% pass rate with the sample size of 8. The students from further education are a very small sample in the full data set. Because of the small sample size of further education node this split may not be relevant. The branches with further education and higher education are not divided further. Because they didn't come directly from school, they may not be influenced by their school academic background. Thus if the students with a lower matriculation aggregate, come from another higher education institution they have 56.1% chance to pass into their second year. Students who have an aggregate between 41-56 and come from school are the worse category (30.9% pass rate). This sample of those who have an aggregate between 41 and 56 and have come from school is now segmented by the taking of Accountancy at school. Not taking Accountancy and standard grade are merged together as there is no significant difference between them and this combined group is separated from higher grade. The completed percentage of the group of students who have an aggregate between 41-56, come from school and do not have Accountancy at higher grade is 25.5% and this sample size is 278; and the group of students who have an aggregate between 41-56, come from school and have Accountancy at the higher grade have a pass rate of 35.8% with the sample size of 302. Thus the students who do not have

Accountancy at higher grade is the worse group (25.5% pass rate for those who did not have Accountancy at higher grade versus 35.8% for those who did). The next meaningful variable for this group is English. The students with an aggregate between 41-56, come directly from school, who do not have Accountancy at the higher grade and English as a first language have 23% pass rate with sample size of 161; those who have English as a second language have a 29.1% pass rate with the sample size of 117. Thus those who have English as a first language group have the lower pass rate. The group with the English as a second language is not divided further. The students with an aggregate between 41-56, come from school, have Accountancy at the higher grade is divided into two groups by age upon starting their degree: Those with age between 17-19 with completed rate of 28.5% and the sample size of 200 and those aged 20-25 have a 50% completed rate and the sample size of 102. The group with an age between 17-19 is worse and cannot be divided further. For the other group of those who have an aggregate between 41-56, come from school, have Accountancy at the higher grade and with an age between 20-25, English is the most significant predictor; the completed rate is 56.2% with the sample size of 73 in the first language and falling to 34.5% with the sample size of 29 in the second language. There are no further significant predictors that divide this group further. Thus, even if the students' get lower aggregate and come from school if they have Accountancy at the higher grade, English as a first language and an age between 20 to 25, they have 56.2% chance of completing the first year successfully. For the group of students who have an aggregate between 41-56, come from school, have Accountancy at the higher grade and English as a first language, age is the next important variable. It is categorised into two subsets those aged 17-19 who have 12.2% pass rate with the sample size of 90 and those aged 20-36 who have 36.6% pass rate with the sample size of 71. These two groups are not divided further.

Accountancy is the next strongest predictor of outcome in students with an aggregate between 57-67. It should be noted that the higher grade and standard grade are grouped together to form a sub group with the sample size of 690 as they are not statistically distinct in students' performance. They are, however, distinct from not taking Accountancy with the sample size of 349. The completed rate decreases from 50.3% in

the subset with Accountancy at the higher grade and standard grade to 30.7% in that with those not taking Accountancy. Previous institution type proved to be the next strongest predictor of outcome in students who have an aggregate between 57-67 and have Accountancy. CHAID splits the previous institution type into two groups; further education and school have been merged as they are not statistically distinct in students' performance from higher education. Those with previous institution type in the subset of further education and school have a 45.2% pass rate with the sample size of 533 and students from the higher education have a pass rate of 67.5% with the sample size of 157. The group with higher education is not divided further. If the students have an aggregate between 57-67 and do not have Accountancy, they are divided two groups by age into 17-19 and 20-35. The students within the age group 17-19 have a complete rate of 20.1% and sample size of 194. Whereas, the age group 20-35 has the pass rate of 43.9% and sample size of 155. These groups are not divided further. For the group of students who have an aggregate between 57-67, have Accountancy and do not come from higher education, the next meaningful variable is age. It is divided into two groups 17-20 with the 44.1% completed rate and the sample size of 492 and 21-24 with 58.5% completed rate and the sample size of 41. These two age groups are not divided further.

The sample is not divided at the child node with the value of the aggregate, 68-73.

Physical Science is the next meaningful variable for the aggregate 74-94. CHAID splits Physical Science into two groups. Standard grade and not taking Physical Science students have completed at a pass rate of 77.7% and the sample size is 103; while those with Physical Science at higher grade students have completed at a rate of 89.9% and the sample size is 298. If the students have an aggregate between 74 to 94 and have Physical Science at the higher grade, then English is the next most important predictor. The other group is not divided further. First language has the 90.7% with the sample size of 291 and the second language has the 57.1% of completed rate with the sample size of 7. Because the sample size 7 is very small, this classification may not be relevant.

Figure 4.1 shows all the partitioning in the analysis (six levels). The data set is not further segmented by CHAID, as no splits of the sixth level segments were statistically significant.

The students in the data can be subdivided into 17 subgroups. The goodness of the segmentation can be evaluated by the comparison of completed rate of the whole sample and the completed rate of the terminal nodes.

Table 4.1 shows all the terminal nodes, completed percentages and students' characteristics. Initially 51.1% of students are completed. The nodes of more than this percentage are highlighted in this table.

The highest completed rate is 90.7% for students who have an aggregate of 74-94, have Physical Science at the higher grade and English as a first language. It should be noted that the students with an aggregate above 68 had completed rate above 55%. The students from higher education also had the same completed rate. The completed rate of the students with lower aggregate marks could be increased by the taking Accountancy high grade and English first language. The students with aggregate marks 41-56 and without Accountancy or Accountancy with standard grade had completed rate below 40%. This aggregate students as English second language also had below 40% completed rate even if they were with Accountancy higher grade. Aggregate category of 57-67 students without Accountancy also had a completed rate below 45% rate.

Table 4.1 Description of terminal nodes

Segment	Number of individuals	Response percentage (%)	Rank order by completed percentage	Students' Characteristics					
				Aggregate	Previous institution type	Accountancy	English	Age	Physical Science
1	90	12.2	17	41-56	School	Not taking or SG	First language	17-19	
2	71	36.6	12	41-56	School	Not taking or SG	First language	20-36	
3	117	29.1	14	41-56	School	Not taking or SG	Second Language		
4	**29**	**56.2**	**7**	**41-56**	**School**	**HG**	**First language**	**20-25**	
5	73	34.5	13	41-56	School	HG	Second language	20-25	
6	200	28.5	15	41-56	School	HG		17-19	
7	**132**	**56.1**	**8**	**41-56**	**HE**				
8	8	37.5	11	41-56	FE				
9	**151**	**67.5**	**3**	**57-67**	**HE**	**HG or SG**			
10	492	44.1	9	57-67	FE or School	HG or SG		17-20	
11	**41**	**58.5**	**5**	**57-67**	**FE or School**	**HG or SG**		**21-24**	
12	194	20.1	16	57-67		Not taking		17-19	
13	155	43.9	10	57-67		Not taking		20-35	
14	**353**	**64.3**	**4**	**68-73**					
15	**291**	**90.7**	**1**	**74-94**			**First Language**		**HG**
16	**7**	**57.1**	**6**	**74-94**			**Second Language**		**HG**
17	**103**	**77.7**	**2**	**74-94**					**Not taking or SG**

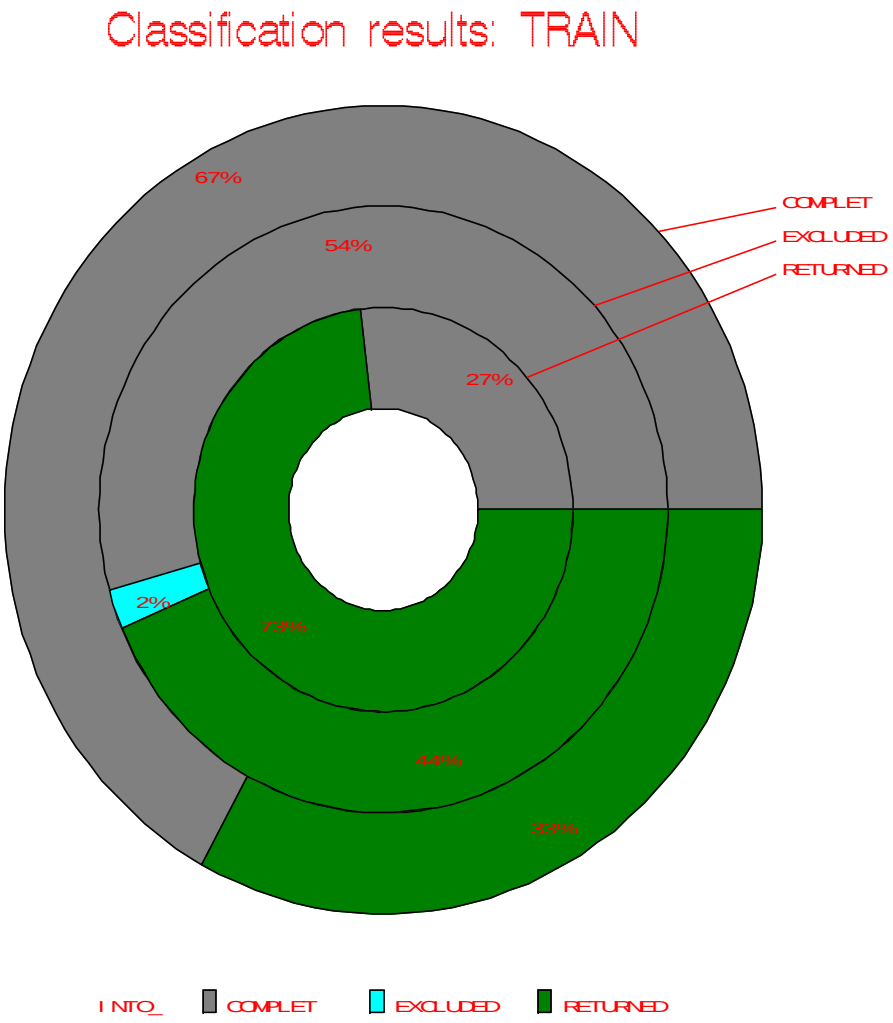

Figure 4.2 Donut chart showing the training data classification summary display generated by using the SAS macro CHAID.

The classification results based on the CHAID analysis are presented in Figure 4.2. The misclassification rates in groups completed, excluded and returned are 33%, 98% and 27%, respectively. Because there were very few students had been excluded in the data set, misclassification was high for the case of excluded. In case of a validation dataset, relatively more classification errors were observed (Figure 4.3). The misclassification rates in groups completed, excluded and returned are 41%, 100% and 36% respectively.

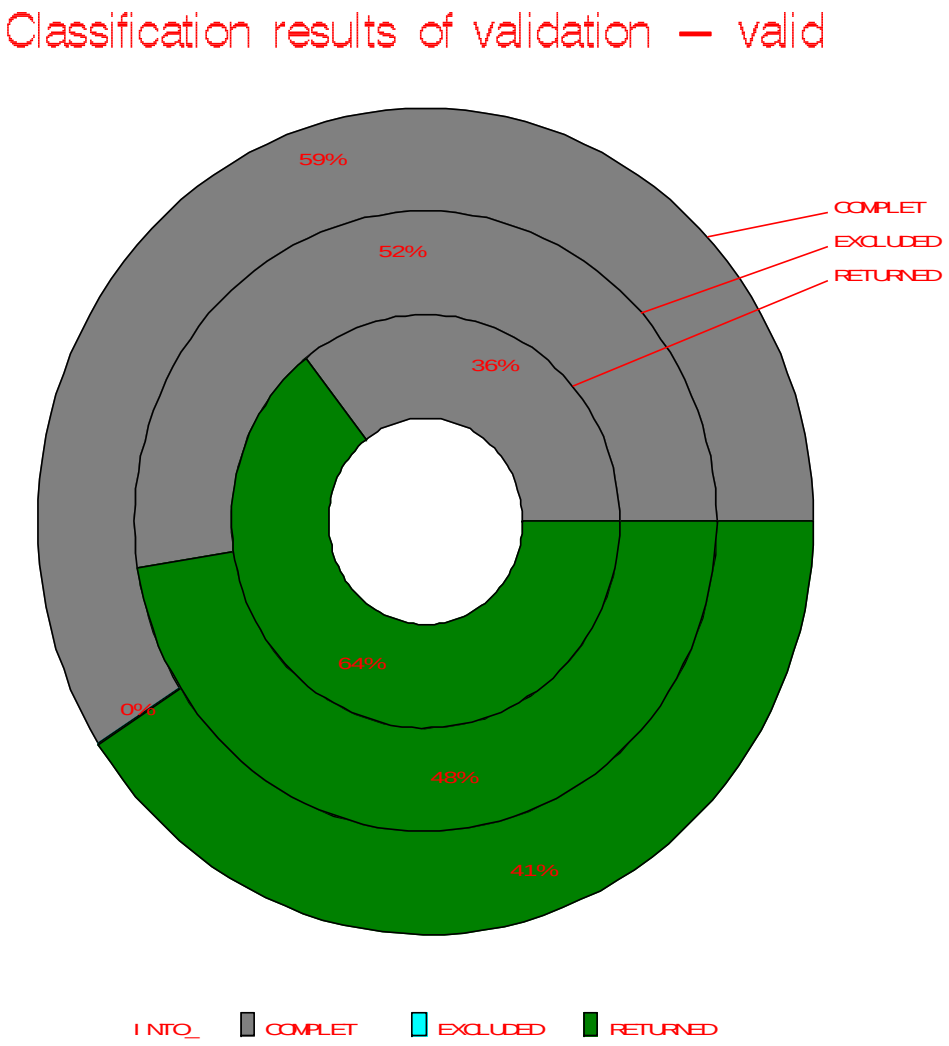

Figure 4.3 Donut chart showing the validation data classification summary display generated by using the SAS macro CHAID.

Observed and predicted percentages are given in the Table 4.2. The training sample which was used to fit the model contains the total of 2511 students, 1284 completed students, 94 excluded students and 1133 returned students. From this, model predicted correctly 861 students as completed, two students as excluded and 828 students as returned.

The validation sample contains the total of 2273 students, 1427 completed students, 63 excluded students and 783 returned students. From this, model predicted correctly 842 students as completed, 505 students as returned.

Overall percentage correct for training and validation data are 67.34% and 59.30% respectively. Predicted accuracies for completed and returned are very high compared to excluded cases as shown in the donut chart. Because excluded students have or have been cancelled their registration from the Faculty for various purposes and this number was very small. As Overall percentage correct greater than 50%, model could be used for prediction.

Table 4.2 Observed and predicted percentages of CHAID.

		Predicted			
Sample	**Observed**	**complete**	**excluded**	**returned**	**Percent Correct**
Training	complete	861	0	423	67.10%
	excluded	51	2	41	2.10%
	returned	305	0	828	73.10%
	Overall Percentage	48.50%	0.00%	51.45%	67.34%
Valid	complete	842	1	584	59.00%
	excluded	33	0	30	0.00%
	returned	278	0	505	64.50%
	Overall Percentage	50.80%	0.00%	49.20%	59.30%

Dependent Variable: Results

The tree was built using the validation data set manually according to the fitted model is shown in Figure 4.4. At the root node the completed rate is 62.8% in the validation data set. It should be noted that the completed percentage trends upward across the aggregate groups as well as in the model tree.

In comparison to the model tree the completed rate of the student who had aggregates of 41- 56 and came from further education in the validation tree is 100%. This is too high as this validation data had only one student in this group. CHAID model tree reveals that

the completed rate of the students from higher education is higher than school. But the validation tree reveals that the completed rate of students from higher education is less than school. It is contradicted as the number of students from higher education is very less than the number of students from school. Otherwise classification theory and completed rates of both trees are similar.

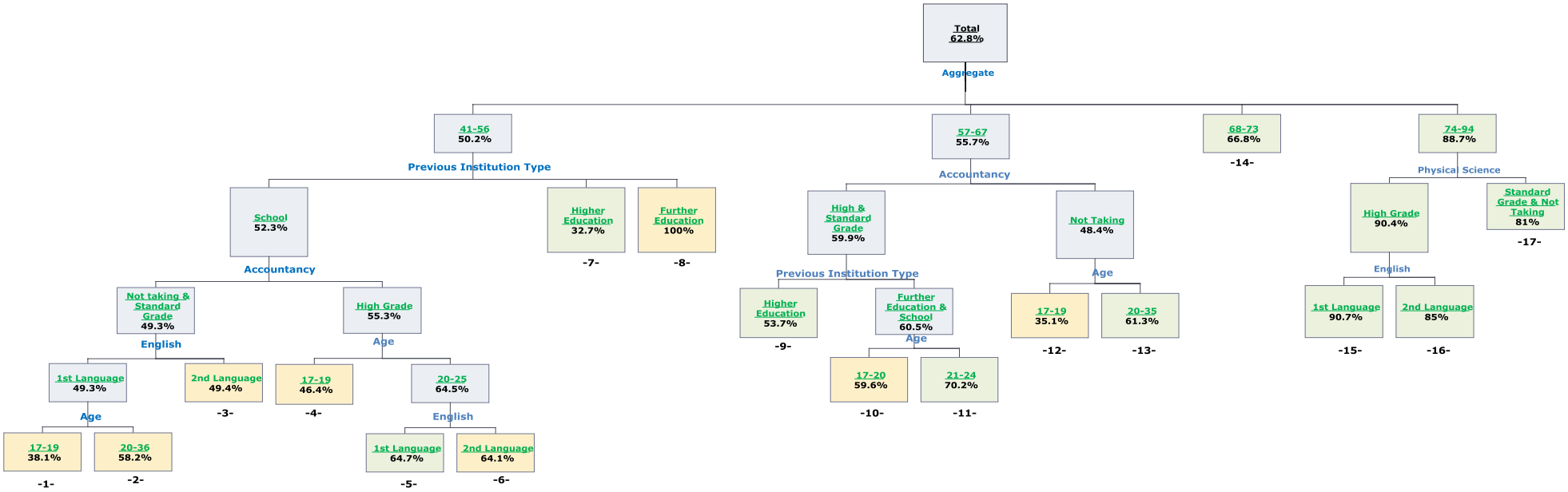

Figure 4.4 Validation tree

4.2 Multinomial Logistic Regression

The multinomial logistic regression model output for the predictor variables is in Table 4.3 and includes the tests for effects of the variables. Except for the last entry, all the chi-square statistics are the Wald statistic, not the likelihood ratio tests statistics. Each chi-square is a test of the null hypothesis that the explanatory variable has no effect on the students' performance.

Table 4.3 Maximum Likelihood Analysis of Variance for the Single variables

Source	DF	Chi-Square	Pr > ChiSq
Intercept	2	86.91	<.0001
Previous Institution Type(ins)	4	46.51	<.0001
Age(age)	2	23.50	<.0001
Aggregate(agg)	2	211.59	<.0001
Accountancy(acc)	4	42.46	<.0001
English(eng)	2	6.69	0.0353
Physical Science(phy)	4	1.03	0.9046
Gender(gen)	2	1.36	0.5074
Biology(bio)	4	6.83	0.1452
History(his)	3*	.	.
Mathematics(mat)	2	3.18	0.2038
Likelihood ratio	4.00E+03	3132.88	1.0000

NOTE: Effects marked with '*' contain one or more redundant or restricted parameters.

The interactions of two factors to six factors are shown in Table 4.4 to Table 4.8 respectively. Significant single predictors and interactions are at a 90% confidence level and are highlighted by bold types in Table 4.3 to 4.8.

The only clear conclusion is that the predictor variables gender and Biology do not contribute significantly to the outcome variable. Because these two variables were not included in any significant predictor or significant two factor interactions. SAS couldn't run for all three factor interactions in the data set. So, Table 4.5 shows the three factor interactions which had been chosen from the significant predictors and two way interactions according to the results in Tables 4.3 and 4.4. Similarly, the possible four, five and six factor interactions were chosen from the three, four and five factor interaction test tables respectively and significant predictors.

Table 4.4 Maximum Likelihood Analysis of Variance for the two factor interaction

Source	DF	Chi-Square	Pr > ChiSq
Intercept	2	12.48	0.0020
age*ins	3*	.	.
ins*gen	4	6.40	0.1710
agg*ins	3*	.	.
ins*acc	7*	.	.
ins*bio	6*	.	.
ins*eng	4	5.24	0.2632
ins*his	4*	.	.
ins*mat	4	3.64	0.4567
ins*phy	4*	.	.
age*gen	2	2.93	0.2309
age*agg	2	9.44	0.0089
age*acc	4	9.56	0.0485
age*bio	3*	.	.
age*eng	2	17.18	0.0002
age*his	4	14.52	0.0058
age*mat	2	0.07	0.9645
age*phy	3*	.	.
age*gen	2	2.96	0.2272
gen*acc	4	1.94	0.7473
gen*bio	4	4.58	0.3337
gen*eng	2	2.20	0.3333
gen*his	2*	.	.
gen*mat	2	1.55	0.4605
gen*phy	4	3.95	0.4125
agg*acc	4	15.12	0.0045
agg*bio	4	3.14	0.5349
agg*eng	2	7.54	0.0231
agg*his	3*	.	.
agg*mat	2	7.91	0.0192
agg*phy	4	9.68	0.0462
acc*bio	8	1.06	0.9978
acc*eng	4	5.18	0.2689
acc*his	5*	.	.
acc*mat	4	1.28	0.8641
acc*phy	8	2.96	0.9371
bio*eng	4	6.18	0.1859
bio*his	7*	.	.
bio*mat	4	3.78	0.4368
bio*phy	8	2.70	0.9517
eng*his	2*	.	.
eng*mat	2	2.19	0.3350
eng*phy	2*	.	.
his*mat	4	3.72	0.4447
his*phy	6*	.	.
mat*phy	4	8.34	0.0800
Likelihood Ratio	4.00E+03	2856.95	1.0000

NOTE: Effects marked with '*' contain one or more redundant or restricted parameters

Table 4.5 Maximum Likelihood Analysis of Three Factor Interactions

Source	DF	Chi-Square	Pr > ChiSq
Intercept	2	128.95	<.0001
age*agg*acc	4	16.14	0.0028
age*agg*eng	2	5.64	0.0597
age*agg*his	4	26.12	<.0001
age*agg*mat	2	8.42	0.0148
age*agg*phy	4	13.98	0.0073
age*acc*his	8	6.94	0.5433
age*eng*his	4	8.60	0.0718
agg*acc*eng	4	8.38	0.0786
agg*acc*mat	4	10.94	0.0273
agg*acc*phy	8	12.54	0.1288
agg*eng*mat	2	6.06	0.0484
agg*eng*phy	4	10.75	0.0296
agg*mat*phy	4	22.49	0.0002
age*mat*phy	4	12.81	0.0122
acc*mat*phy	8	13.78	0.0876
eng*mat*phy	4	2.54	0.6374
Likelihood Ratio	3.00E+03	2154.43	1.0000

Table 4.6 Maximum Likelihood Analysis of Four Factor Interactions

Source	DF	Chi-Square	Pr > ChiSq
Intercept	2	135.79	<.0001
age*agg*eng*acc	4	7.45	0.1140
age*agg*acc*his	8	66.55	<.0001
age*agg*eng*his	4	10.44	0.0336
age*agg*eng*mat	2	11.45	0.0033
age*agg*eng*phy	4	11.94	0.0178
age*agg*acc*mat	4	45.57	<.0001
age*acc*eng*his	8	23.97	0.0023
agg*eng*acc*mat	4	47.42	<.0001
agg*acc*mat*phy	8	38.93	<.0001
age*agg*acc*phy	8	67.56	<.0001
agg*eng*acc*phy	8	71.55	<.0001
age*agg*mat*phy	4	65.47	<.0001
agg*eng*mat*phy	4	7.33	0.1193
age*acc*mat*phy	8	24.19	0.0021
age*eng*mat*phy	4	8.34	0.0799
eng*acc*mat*phy	8	50.82	<.0001
Likelihood Ratio	3.00E+03	2162.26	1.0000

Table 4.7 Maximum Likelihood Analysis of Variance for five factor Interactions

Source	DF	Chi-Square	Pr > ChiSq
Intercept	2	202.78	<.0001
age*agg*acc*eng*mat	4	20.98	0.0003
age*agg*acc*mat*phy	8	29.91	0.0002
agg*acc*eng*mat*phy	8	54.15	<.0001
age*agg*acc*eng*phy	8	3.06	0.9307
age*agg*acc*his*eng	8	13.69	0.0901
age*agg*eng*mat*phy	4	9.89	0.0423
age*acc*eng*mat*phy	8	57.35	<.0001
Likelihood Ratio	3.00E+03	2426.32	0.9998

Table 4.8 Maximum Likelihood Analysis of Variance for Six Factor Interaction

Source	DF	Chi-Square	Pr > ChiSq
Intercept	2	459.96	<.0001
age*agg*acc*mat*eng*phy	8	115.00	<.0001
Likelihood Ratio	3.00E+03	2461.53	1.0000

Significant predictors and interactions from the above tables were put together and the tests of those significant predictors and interactions are given in Table 4.9. Significant variables are highlighted by bold types. Examination of the Wald statistics in Table 4.9 suggests that:

- Age;
- English;
- age*aggregate*Accountancy;
- age*aggregate*Mathematics;
- age*Accountancy*English;
- aggregate*English*Mathematics;
- aggregate*Mathematics*Physical Science;
- age*aggregate*English*Physical Science;
- age*aggregate*Mathematics*Physical Science;
- age*aggregate*Accountancy*English*Mathematics;
- age*aggregate*Accountancy*English*Mathematics*Physical Science;

may contribute to the model.

The model was fitted using the significant variables in Table 4.9 and the tests of those results are shown in Table 4.10. Table 4.10 shows the global tests for the effect of those variables on the students' performance. English, age*Accountancy*English and

age*aggregate*Accountancy*English*Mathematics are not significant variables; the other variables in the table are significant. The final model was fitted using the significant variables of Table 4.10, coding of the variables were examined and the results are shown in Table 4.11.

Table 4.9 Maximum Likelihood Analysis of Variance

Source	DF	Chi-Square	Pr > ChiSq
Intercept	2	13.86	0.0010
Age	2	11.31	0.0035
Aggregate	1*	.	.
Accountancy	3*	.	.
English	2	5.85	0.0538
age*agg	1*	.	.
age*acc	3*	.	.
age*eng	1*	.	.
agg*acc	3*	.	.
agg*eng	1*	.	.
agg*mat	1*	.	.
agg*phy	3*	.	.
mat*phy	3*	.	.
age*agg*acc	4	23.17	0.0001
age*agg*eng	1*	.	.
age*agg*mat	2	5.38	0.0677
age*agg*phy	2*	.	.
age*acc*eng	4	9.03	0.0603
agg*acc*eng	3*	.	.
agg*acc*mat	3*	.	.
agg*eng*mat	2	5.13	0.0771
agg*eng*phy	3*	.	.
agg*mat*phy	4	24.94	<.0001
acc*mat*phy	7*	.	.
age*mat*phy	2*	.	.
age*agg*acc*mat	2*	.	.
age*agg*eng*mat	0*	.	.
age*agg*eng*phy	4	9.55	0.0488
age*acc*mat*phy	5*	.	.
age*eng*mat*phy	4	6.85	0.1442
agg*acc*eng*mat	3*	.	.
agg*acc*mat*phy	5*	.	.
age*agg*acc*phy	6*	.	.
agg*acc*eng*phy	5*	.	.
acc*eng*mat*phy	6*	.	.
age*agg*mat*phy	4	17.57	0.0015
age*agg*acc*eng*mat	4	14.78	0.0052
age*agg*acc*mat*phy	5*	.	.
agg*acc*eng*mat*phy	7*	.	.
age*acc*eng*mat*phy	6*	.	.
age*agg*acc*eng*mat*phy	8	17.16	0.0285
Likelihood Ratio	2.00E+03	1804.26	1.0000

NOTE: Effects marked with '*' contain one or more redundant or restricted parameters

Table 4.10 Maximum Likelihood Analysis of Variance

Source	DF	Chi-Square	Pr > ChiSq
Intercept	2	9.77	0.0076
Age	2	5.18	0.0750
English	2	3.59	0.1661
age*agg*acc	4	51.40	<.0001
age*phy*his	8	17.25	0.0276
age*agg*mat	2	25.10	<.0001
age*acc*eng	4	7.68	0.1041
agg*eng*mat	2	11.02	0.0041
age*agg*acc*eng*mat	4	3.78	0.4431
age*agg*acc*eng*mat*phy	8	25.90	0.0011
Likelihood Ratio	3.00E+03	2514.61	1.0000

Table 4.11 Estimated coefficients, estimated standard errors, Wald statistics and two tailed p-values for the full multivariable model fit

Logit	Variable	Coeff.	Std.Err.	Chi-Square	Pr>ChiSq
1	**Intercept**	-2.0060	0.63400	10.00	0.0016
	Age	0.0780	0.03200	5.82	0.0158
	age*agg*acc(1)	0.0005	0.00010	55.76	<.0001
	age*agg*acc(2)	-0.0001	0.00010	1.37	0.2417
	age*agg*mat(1)	0.0002	0.00005	16.70	<.0001
	agg*eng(1)*mat(1)	0.0040	0.00100	15.33	<.0001
	age*agg*acc(1)*eng(1)*mat(1)*phy(1)	0.0001	0.00010	1.76	0.1840
	age*agg*acc(1)*eng(1)*mat(1)*phy(2)	0.0001	0.00010	1.09	0.2956
	age*agg*acc(2)*eng(1)*mat(1)*phy(1)	0.0002	0.00010	7.50	0.0062
	age*agg*acc(2)*eng(1)*mat(1)*phy(2)	-0.0002	0.00010	5.04	0.0248
2	**Intercept**	-2.71200	1.4760	3.37	0.0662
	Age	-0.00800	0.0760	0.01	0.9160
	age*agg*acc(1)	0.00004	0.0002	0.08	0.7776
	age*agg*acc(2)	0.00020	0.0002	2.00	0.1575
	age*agg*mat(1)	0.00040	0.0001	10.47	0.0012
	agg*eng(1)*mat(1)	0.00200	0.0030	0.36	0.5463
	age*agg*acc(1)*eng(1)*mat(1)*phy(1)	-0.00010	0.0002	0.42	0.5167
	age*agg*acc(1)*eng(1)*mat(1)*phy(2)	0.00020	0.0002	0.83	0.3624
	age*agg*acc(2)*eng(1)*mat(1)*phy(1)	-0.00010	0.0002	0.42	0.5150
	age*agg*acc(2)*eng(1)*mat(1)*phy(2)	0.00020	0.0002	0.96	0.3269
	Likelihood Ratio			2277.78	0.9836

Logit 1 indicates the model for completed versus returned and logit 2 indicates the model for excluded versus returned. The likelihood ratio test is a goodness-of-fit test. High p-

values indicate adequate fit. The chi-square value of likelihood ratio of the model in Table 4.11 is greater than that shown in Table 4.10. So the model in Table 4.11 is a better fit than that in Table 4.10.

The Chi-square value of the model in Table 4.11 is 2277.78 and the corresponding p-value is 0.9836, implying the model fits the data adequately. The effects of the interaction age*aggregate*Mathematics on the log odds of both logits are highly significant at 10% significance level.

The six factor interactions age*aggregate*(Accountancy-not taking versus standard grade) *(English-first language versus second language)*(Mathematics-high grade versus standard grade)*(Physical Science-high grade versus standard grade) contribute positively and most significantly on the model completed versus returned. The other six factor interaction age*aggregate*(Accountancy-not taking versus standard grade) *(English-first language versus second language)*(Mathematics-high grade versus standard grade)*(Physical Science-not taking versus standard grade) contribute negatively on that model.

The three factor interactions age*aggregate*(Accountancy higher grade versus standard grade), age*aggregate*(Mathematics higher grade versus standard grade) and aggregate* (English first language versus second language)* (Mathematics higher grade versus standard grade) influence positively on the completed versus returned model most significantly. “These three, three-factor” interactions are the most significant variables in this model.

Age is the only significant single variable on the completed versus returned model. On this model, for one unit change in the variable age, the log of the ratio of the two probabilities, P(completed)/P(returned), will be increase by 1.08. Therefore when a student’s age increases, he/she will have a better probability to complete the first year.

The three factor interaction age*aggregate*(Mathematics-higher grade versus standard grade) is the only significant variable in the excluded versus returned model.

Then this model was applied on the validation data and the predicted frequencies are given in the Table 4.12. The validation sample contains the total of 2274 students, 1428 completed students, 63 excluded students and 783 returned students. From this, model predicted correctly 948 students as completed, 573 students as returned.

The Overall percentage predicted correctly is 66.91%. Prediction is good for completed and returned cases. Predicted accuracies for completed and returned are very high compare to the excluded group. Because the excluded students have been cancelled by Faculty for various reasons and this number was very small. As Overall percentage correct greater than 50%, model may be used for prediction.

Table 4.12 Predicted Frequencies of the multinomial logistic regression model

	Predicted			
Observed	**Completed**	**Excluded**	**Returned**	**Percent Correct**
Completed	948	1	479	66.39%
Excluded	37	0	26	0%
Returned	210	0	573	73.24%
Overall Percentage	52.55%	0.0%	47.41%	66.91%

CHAPTER 5

DISCUSSION AND CONCLUSIONS

This report used data from the Faculty of Commerce, University of the Witwatersrand from the four cohorts of 2003 to 2006 to examine the performance of the first year students. This final chapter presents the discussions of the results of this report. It also presents a brief overview of the study and answers to the research questions which drove this study and connects them to the relevant literature and theory. This chapter will conclude with a discussion of the study's limitations and implications for future research and practice.

Exclusion rates for first year students in the Faculty of Commerce were 38%, 40% and 38% in the years 2000, 2001 and 2002 respectively; A further 15%, 16% and 12%, respectively have been required to repeat the first year (Oracle system by Management Information Unit at University of the Witwatersrand, 2007).

The goal of this report is to find the important predictors for first year students' performance in the Faculty of Commerce, University of the Witwatersrand. Operationally, the following questions to be answered by the study were formulated:

- Does matriculation Aggregate, some common matriculation courses (Accountancy, Biology, English, History, Mathematics and Physical Science), Gender, Age on enrolment and previous Institution type predict first year performance?
- Which variables, at what stages, are most efficient in predicting "completed", that is completed first year successfully?

- How do the different factors combine and interact with each other in predicting first year success?

5.1 Discussion on the results of the CHAID analyses

As discussed in the literature review, CHAID and multinomial logistic regression are useful for identifying possible important predictors of this problem, when the predicted variable is polychotomous. Both analyses require large sample sizes in order to obtain reliable and replicable results. These results and analyses demonstrate that CHAID and multinomial logistic regression can be used effectively with education data. CHAID has been used to analyse different performance groupings of MBA students based on various entry (selection) characteristics (Perreault & Wagner, 1979). The trend in higher education is for researchers to recognise the limitations of ordinary least squares regression and turn increasingly to logistic regression for explaining relationships between a categorical outcome variable and a mixture of continuous and categorical predictor variables. This trend is primarily motivated by complex data and categorical outcome measures, for example, enrolment, retention and graduation that are of interest to higher education researchers. It has facilitated research on this topic by Hinkle, Austin and McLaughlin(1988), Austin, Yaffe and Hinkle(1992), Dey and Astin(1993) and Cabrera(1994) by using logistic regression as an analytic tool. For this research, predictors such as previous Institution type, Age, Gender, matriculation Aggregate and the study of some of the common matriculation subjects (Accountancy, Biology, English, History, Mathematics and Physical Science) are used. This research examined the relationship between those variables and the outcome of the student's first year.

The data was partitioned into two data sets: A training data set (the 2003 & 2004 cohorts) and a validation data set (the 2005 & 2006 cohorts). The training data was used to estimate the model and the validation data was used to test the fitted model by assessing the misclassification rate. In the training data set, the completed rate was 51. 1% whilst in the validation data set the completed rate was 62.8%.

From the CHAID analysis, the most salient dimension for students' performance in these two data sets is the matriculation aggregate. Previous institution type, Accountancy, Physical Science, English and age have significant influence in the prediction of students' performance by interacting with aggregate and some of these variables in different orders with differing success rates. In this study Gender, Biology, History and Mathematics have shown no significant influence in the prediction of students' performance in the model when using CHAID. This analysis shows that the completed percentage of first year trends upward across the groups of aggregate, from lowest to highest. Students admitted with a matriculation aggregate of 74-94% have approximately a 2.5 times better completed rate when compared with those with a matriculation aggregate of 41-56%.

The results of the CHAID show that the best subgroup's completed rate is 90.7% with a total sample of 401 students who have an aggregate of 74-94%, have Physical Science at the higher grade and English as a first language.

The students with an aggregate above 68% have a completed rate above 55% in all subgroups: i.e. the subgroup of students with:

(i) an aggregate between 68-73; and

(ii) the subgroup of students who have an aggregate between 74-94, English as first language and Physical Science at higher grade:

(iii) the subgroup of students who have an aggregate between 74-94, English as second language and Physical Science at higher grade; and

(iv) the subgroup of students who have an aggregate between 74-94 and do not take Physical Science or have Physical Science at standard grade.

There is one further subgroup of students who also have a completed rate above 55% with a total sample of 132, and these are the students who have an aggregate between 41% and 56% and have come from higher education.

The CHAID results, based on the training set, presented in Chapter 4 consistently confirm the completed rate of the students with lower aggregate marks are increased by taking Accountancy at higher grade and English as a first language.

It has also been found that three groups of students with common characteristics, matriculation aggregate of 41-56 and Accountancy not at higher grade have a completed rate below 40%. Those groups are,

- The subgroup of students who have an aggregate between 41-56, come from school directly, do not take Accountancy or have Accountancy at standard grade, English as first language and aged between 17-19 with completed rate 12.2%,
- The subgroup of students who have an aggregate between 41-56, come from school directly, do not take Accountancy or Accountancy at standard grade, English as first language and aged between 20-36 with completed rate 36.6% and
- The subgroup of students who have an aggregate between 41-56, come from school directly, do not take Accountancy or Accountancy at standard grade, English as second language with completed rate 29.1%.

Thus, if the lowest aggregate students do not take Accountancy or do not take Accountancy at higher grade, they have lower chance in completing their first year successfully.

Results showed that the subgroup of students who have an aggregate between 41-56, come from school directly, had Accountancy at higher grade, English as second language and aged between 20-25, also have a completed rate below 40%. But the students who have these same characteristics but having English as a first language have a completed rate of 56.2%. Thus, even if the lowest aggregate students aged between 20-25 and have Accountancy at higher grade, then taking English as first language increases the completed rate.

The following two groups of students with common characteristics of an aggregate of 57-67 and not taking Accountancy have a completed rate below 45%:

- The subgroup of students who have an aggregate between 57-67, do not take Accountancy and are aged between 17-19 have a completed rate of 20.1% and
- The subgroup of students who have an aggregate between 57-67, do not take Accountancy and are aged between 20-35 have a completed rate of 43.9%.

Thus, if the moderate aggregate students also who come from any institution, do not take Accountancy they also have a poor completed rate.

The CHAID model was validated by applying the classification criteria to the independent validation data set and verifying the success of classification. Because excluded students have had their registration cancelled by the Faculty for various reasons and this number was very small. As the number in this category is small, a few misclassification leads to a high misclassification rate. Thus misclassification rate is high for the case of excluded only in model based on the CHAID analysis. But the predicted accuracies are good for both the completed and returned categories. However, the overall percentage correct for training and validation data are 67.34% and 59.30% respectively. As the overall percentage correct is greater than 50%, the model could possibly be used for prediction.

The validation data were classified manually according to the fitted classification CHAID model and the validation tree was drawn. It was built by grouping the validation data and the completed percentage was calculated for every node. The completed percentages of every node were compared with the fitted classification tree. For every splitting predictor, which category of that predictor has more completed percentage were compared as the overall completed rate of training and validation data differs. Both the validation tree and CHAID model tree are similar except for the case of the previous institution type. When the previous institution type is the splitting predictor, the students from higher education have a better completed rate than those directly from school in the fitted CHAID model. However in the validation tree, the students from higher education have a lower completed rate than those from school. As the number of students from higher education is small, this conclusion is open to question.

5.2 Discussion on the Multinomial Logistic Regression analyses

In the Multinomial Logistic Regression Analysis, Age, Aggregate, Accountancy, English, Mathematics and Physical Science are the significant predictors. Most of these variables were significant as variables interacting with some of these variables. Age is the only single variable significant on its own in these models. These models show that when a student's age increases, it is more likely that he/she will complete the first year successfully. The interactions of the predictor variables age, aggregate, Accountancy, Mathematics, English and Physical Science are:

(i) age, aggregate, and Accountancy-higher grade versus standard grade;

(ii) age, aggregate, and Mathematics-higher grade versus standard grade;

(iii) aggregate, English-first language versus second language, Mathematics-higher grade versus standard grade;

(iv) age, aggregate, Accountancy-not taking versus standard grade, English-first language versus second language, Mathematics-higher grade versus standard grade, and Physical Science-not taking versus standard grade;

(v) age, aggregate, Accountancy-not taking versus standard grade, English-first language versus second language, Mathematics-higher grade versus standard grade, Physical Science-higher grade versus standard grade;

all contribute significantly in the Multinomial Logistic Regression model.

Multinomial Logistic Regression was validated using goodness-of-fit=statistics. The Chi-square (degrees of freedom = 60) value of the model is 2277.78 and the corresponding p-value is 0.9836, implying the model fits the data adequately. The fitted multinomial logistic regression model using the training data was also applied to the validation data set to validate the model. The observed and predicted frequencies were calculated and overall error rate was determined. The Overall percentage predicted correctly is 66.91%. The prediction rate for completed and returned students is good. The predicted accuracies

for completed and returned students are very high compared to the excluded group, because the excluded students have been cancelled by Faculty for various reasons and this number was also very small. As the overall percentage correct is greater than 50%, the model may be used for prediction.

5.3 General discussions on the models

Aggregate, Age, Accountancy and English have been found to be important predictors of the students' performance as a single variable or in interactions in both analyses, CHAID and multinomial logistic regression. Some of the significant interaction variables contain the previous institution type in CHAID but not in the Multinomial Logistic Regression analyses. Some of the significant interaction variables contain Mathematics in the multinomial logistic regression but not in the CHAID analyses. Gender, Biology and History have no significant influence on students' performance as single or as significant interaction variables as the results of both analyses did not contain these variables. Aggregate is the only significant single predictor in CHAID. Age is the only significant single predictor in multinomial logistic regression analysis. As the major single variable differs in both models, the probability of the prediction of the students' performance also will differ. However, either one of the two models can be used for prediction. Because the accuracy of prediction of both models is good.

The accuracy of prediction is good for the completed and returned categories. It is a common feature of both models. As has been discussed, the multinomial logistic regression and CHAID models are fairly similar overall. CHAID is simpler to interpret when there are a lot of categorical predictors. The accuracy of prediction of the multinomial logistic regression is better than that of CHAID. But both CHAID and multinomial logistic regression are similar in predicting performance.

5.4 Important categories of the significant variables

The students who have an aggregate between 74-94 have the highest completed rate compared to all the other aggregate groups in the first split in the CHAID analyses.
When previous institution type is the splitting variable for the subgroup of students who have an aggregate between 41-56 and the subgroup of students who have an aggregate between 57-67 and Accountancy, those who come from higher education have higher completed rate than those from further education and school.

When Accountancy is the splitting variable for the subgroup of students who have an aggregate between 57-67, the students who took Accountancy have a higher completed rate than those who did not take Accountancy.

When Physical Science is the splitting predictor for the group of students who have an aggregate between 74-94, higher grade is the more advantageous in increasing the completed rate.

The students who have English as first language have a higher completed rate than those who have English as a second language when English is the splitting variable for the following two groups of students:

- The subgroup of students who have an aggregate between 41-56, come from school directly, Accountancy at higher grade and aged 20-25 and
- The group of students who have an aggregate between 74-94 and Physical Science at higher grade.

The students aged older than 20 years have a higher completed rate than those younger than 20 years when age is the splitting variable for the following groups of students:

- The subgroup of students who have an aggregate between 41-56, come from school, do not take Accountancy or Accountancy at standard grade and English as first language,
- The group of students who have an aggregate between 41-56, come from school directly and Accountancy at higher grade,
- The group of students who have an aggregate between 57-67, take Accountancy and come from school or further education and
- The group of students who have an aggregate between 57-67 and do not take Accountancy.

In the logistic regression an interaction that contains Mathematics at higher grade has a greater chance of successful completion of the first year than those that have standard grade Mathematics.

Thus, from these groupings, we can conclude that the categories 74-94 in aggregate, higher education in previous institution type, taking Accountancy in Accountancy, higher grade in Physical Science, first language in English, older than 20 years in age and higher grade in Mathematics are positive in increasing the completed rate.

This results of this research show that a student's gender did not affect the student's performance during the first year of University in the Faculty of Commerce. This finding is consistent with prior research: Astin (1997) had found that gender only explains 2% of the variance in a retention study. But, Lembesis (1965) found that women tended to get better grades than the men during the first year, but that study focused exclusively on students who have or have been cancelled their registration.

The matriculation mark is a reasonably good predictor of pass/fail at University (Mitchell et al, 1997). Robbins (2004) found that approximately 25 percent of the variance in the students' performance can be attributed to their high school performance. Consistent with prior research this study also found that upon entering university, past academic ability

played an important role in the first year of University; in fact a student's high school aggregate is the most influencing variable of first year performance from the CHAID analysis.

Students admitted with an aggregate of 74-94 have approximately 2.5 times improvement in completed rate when compared with those with an aggregate of 41-56 when the other predictors were not considered. This finding is consistent with the prior research that it seems that only a small percentage of those students with a school result of below 70% obtain a first-year university average performance of 50% or more (Roux, Bothma & Botha, 2004).

Consistent with existence research, school marks for Mathematics, Science and English were all related to first year performance (Eeden, Beer & Coetzee, 2001).

The results obtained for the two statistical techniques are similar but not identical. The entry requirements to improve the current percentage of completed were shown in the Table 4.1. The Multinomial Logistic Regression fitted model is described in Chapter 4.2. A student's completed chance can be calculated using it from his/her entering characteristics.

The study is aimed at exploring some of the important predictors of students' performance in first year in the Faculty of Commerce. This was achieved by applying both CHAID and multinomial logistic regression methodologies. Based on the findings of this study, attention should be paid for the students' academic performances at high school importantly matriculation Aggregate, Accountancy, English, Mathematics and Physical Science. Previous Institution type may also contribute to the students' performances. Age is also vital to the students' academic performances.

An important limitation of this study is that all data were drawn from Faculty of Commerce, University of the Witwatersrand, which limits the ability to generalise these

findings. One can be confident in knowing that some of the matriculation courses and aggregate are truly influencing first year performance. But caution is warranted if attempting to generalise these results to other Universities unless these universities have students with similar characteristics. This model can not be used for other Faculties also in the University of the Witwatersrand as the admission criteria and requirements are different from the Faculty of Commerce. This is another limitation of this study.

Data were collected from the four cohorts 2003 to 2006 in this study. The cohorts 2003 and 2004 were used as training data to fit the model and 2005 and 2006 were used as validation data to validate the fitted model. But the training data and validation data sets have different pass rates. In the training data set, the completed rate was 51. 1% whilst in the validation data set the completed rate was 62.8%.This is a further limitation. This may also have affected the prediction of completed rate.

A further limitation to this study is that many variables previously shown to influence students' performance were not gathered and included in the analysis. This limitation was born of a practical reality; many of these variables were not readily available. For example, programmatic structures, race, family income, financial aid, participation in university sports, on or off-campus jobs etc. were not and could not be included in this study. Perhaps the most important variable which were not included in this study were direct measurements of commitment to university

It is a start of the research. Findings of this study will help as the basis for further research into higher education. A large number of questions concerning first year students' performance in the Faculty of Commerce, University of the Witwatersrand, require further attention. Among these, it is suggested that future research look into the areas of inquiry such as how early the student starts reading for exams, fathers' education, income category of the father, family background, type of the high school, etc and predictors in this study also and analyse this data using mixture models of both a quantitative and qualitative studies. Most of the variables which were described in the

previous paragraph are qualitative variables and have to be collected directly from the students. This information could be collected by preparing the appropriate questionnaire. It may also be useful to pursue a robust programme of qualitative and quantitative studies.

By selecting first year Commerce students using these criteria it is likely that first year performance will be improved.

REFERENCES

Aitken, D. N. 1982. College student performance, satisfaction and retention: Specification and estimation of a structural model. *The Journal of Higher Education*. 53, 32-50

Agresti, A. 1990. *Categorical Data Analysis.* Wiley Series, New York.

Albert, A. & Anderson, J. A. 1984. On the existence of maximum likelihood estimates in logistic regression models. *Biometrika*. 71, 1-10.

Aldrich, J. H. & Nelson, F. D. 1984. *Linear probability, probit and logit models. SAGE*, Beverly Hills and London.

Alexander, N. 2001. Language policy, symbolic power and democratic responsibility of the post-apartheid University. *Paper presented at D.C.S. Ooisthuizen memorial lecture*. Rhodes University, Grahamstown, 9 October.

Amoore, H. 2001. The Matriculation Board. An account of the position of matriculation in 2001. In *The challenges of access and admissions (p.27-38),* Pretoria: SAUVCA.

Arentze, T. & Timmermans, H. 2003. Measuring the goodness-of-fit of decision-tree models of discrete and continuous activity-travel choice: Methods and empirical illustration. *Journal of Geographical Systems,* 5, 185-206.

Astin, A. 1971. Open admissions and programs for the disadvantaged. *The Journal of Higher Education*. 42, 629-647.

Astin, A. 1972. College dropouts: A national profile. *ACE Research Reports*, 7, Washington, D.C.: American Council on Education.

Astin, A. 1997. How "Good" is your Institution's retention rate?. *Research in Higher Education*, 38, 647-658.

Austin, J. T., Yaffee, R. A. & Hinkle, D. E. 1992. Logistic regression for research in higher education. *Higher Education: Handbook of Theory and Research.* 8, 379-410.

Babinec, T. 1990. CHAID response modeling and segmentation. *Quirck's Marketing Research Review*. June, 12-15.

Bargate, K. 1999. Mathematics as an indicator of success in first year Accounting programmes at Technikon, Natal. *South African Journal of Higher Education*. 13, 139-143.

Barndorff-Nielsen, O. 1978. *Information and exponential families in statistical theory*. New York: J. Wiley and Sons.

Braxton, J. M., Milem, J. F. & Sullivan, A. S. 2000. The influence of active learning on the College student departure process. *The Journal of Higher Education*. 71, 569-590.

Breimen, L. 1996. Bagging predictors. *Machine Learning*, 24, 123-140.

Bult,J. R. & Wansbeek,T. 1995. Optimal selection for direct mail. *Marketing Science.* 14, 378-394.

Cabrera, A. F. 1994. Logistic regression analysis in higher education: An applied perspective. *Higher Education: Handbook of Theory and Research.* 10, 225-256.

Chan,Y. H. 2004. Logistic regression analysis, *Singapore Medial Journa.,* 45, 149-153.

Chaturvedi, A. & Green, P. E. 1995. SPSS for Windows, CHAID 6.0. *Journal of Marketing Research.* 32, 245-254.

Chikte, U. M. E & Brand, A. A. 1996. Diversity in South African Dental Schools. *Journal of the Dental Association of South Africa*. 51, 641-646.

Cope, R. 1971. An investigation of entrance characteristics related to types of College dropouts. *Washington, D.C.:Office of Education Reports*.

Cox, D. R. 1970. *The Analysis of Binary Data*. London: Methuen and Co.

Delvare, I. 1995. Tertiary pass rates in South Africa. *South African Institute of Race Relations.*

Dey, E. L. & Astin, A. W. 1993. Statistical alternatives for studying College student retention: A comparative analysis of logit, probit, and linear regression. *Research in Higher Education*. 34, 569-581.

Diepen, M. V. & Franses, P. H. 2006, Evaluating Chi-squared automatic interaction detection, *Information Systems.* 31, 814-831.

Dyer, H. 1968. School factors and equal educational opportunity. *Harvard Educational Review*, 37, 38-56.

Eeden, V. R., Beer, D. & Coetzee, C. H. 2001. Cognitive ability, learning potential, and personality traits as predictors of academic achievement by engineering and other science and technology students. *South African Journal of Higher Education*. 15, 171-179.

Eherler,D. & Lehmann, T. 2001. Responder profiling with CHAID and dependency analysis. http://www.informatik.uni-freiburg.de/~ml/ecmlpkdd/WS-Proceedings/w10/lehmann.pdf.

Fernandez, G. 2003. Data mining using SAS applications. *Chapman & Hall/CRC*, USA.

Foxcroft, C. & Stumpf, R. 2005. What is matric for? *Paper presented at the CHET-UMALUSI Semina*r, Pretoria. 23 June 2005.

Goduka, I. N. 1996. "Challenges to traditional White Universities: Affirming diversity in the curriculum. *South African Journal of Higher Education*. 10, 27-39.

Haughton, D. & Oulabi, S. 1997. Direct marketing modeling with CART and CHAID. *Journal of Direct Marketing.* 11, 42-52.

Hinkle, D. E., McLaughlin, G. W. & Austin, J. T. 1989. Using log-linear models in higher education research. *New Directions for Institutional Research.* 58, 23-41.

Hoare, R. 2004. Using CHAID for classification problems. *Presented at the New Zealand Statistical Association 2004 conference*. Wellington.

Hosmer, D. W. & Lemeshow, S. 2000. Applied Logistic Regression. *Wiley-Interscience*, USA.

http://www.wits.ac.za.

Huysamen, G. K. 1999. Psychometric explanations for poor predictability of the tertiary academic performance of educationally disadvantaged students. *South African Journal of Higher Education*. 11, 65-71.

Huysamen, G. K. 2001. Marking standards and differential predictability of the first-year University performance of different demographic groups. *South African Journal of Higher Education*. 15, 129-137.

Ian, S., Nan, Y. & Jane, H. 2007. A case for improving teaching and learning in South African higher education. *Research paper prepared for the Council on higher education.*

Ishler, J. L. C. & Upcraft, M. L. 2005. In M. L. Upcraft, J. N. Gardener & B. O. Barefoot (Eds.) *Challenging and supporting the first-year student: A handbook of improving the first year of College*. San Francisco: Jossey-Bass.

Jawitz, J. 1995. Performance in first- and second-year engineering at UCT. *South African Journal of Higher Education.* 9, 101-108.

Jonker, J., Franses, P. H. & Piersma, N. 2002. Evaluating direct marketing campaigns; Recent findings and future research topics. *ERIM Reports ERS-2002-26-MKT*, Erasmus University, Rotterdam.

Kass, G. V. 1975. Significance testing in and some extensions of Automatic Interaction Detection. *Unpublished Ph.D. Thesis*, University of Witwatersrand, South Africa.

Kass, G. V. 1975. Significance testing in Automatic Interaction Detection. *Applied Statistics,* 24, 178-189.

Kass, G. V. 1980. An exploratory technique for investigating large quantities of categorical data. *Applied Statistics*. 29, 119-127.

Kendall,M. G. & Stuart, A. 1961. *The advanced theory of statistics*. London:Griffin.

Lehmann, E. L. 1959. *Testing statistical hypotheses*. New York: J. Wiley and Sons.

Lembesis, A. 1965. A study of students who withdrew from college during their second, third or forth years. *Unpublished Doctoral Dissertation*, University of Oregon.

Levin, N. & Zahavi,J. 2001. Predictive modeling using segmentation. *Journal of Interactive Marketing*. 15, 2-22.

Lolwana, P. 2004. The history of falling Matric standards. *Paper submitted for publication by the HSRC.*

Long, J. S. 1997. *Regression models for categorical and limited dependent variables*. SAGE, Beverly Hills and London.

Mail & Guardian. 2006. Shock Varsity dropout stats. *Mail & Guardian*. 22-28 September.

McFadden, D. 1974. The measurement of urban travel demand. *Journal of Public Economics.* 3, 303-328.

McKenzie, K. & Schweitzer, R. 2001. Who succeeds at the University? Factors predicting academic performance in first year Australian University students. *Higher Education Research and Development*. 20, 21-33.

Menard, S. 1995. A developmental test of Mertonian Anomie theory. *Journal of Research in Crime and Delinquency*. 32, 136-174.

Mitchell, G., Fridjhon, P. and Haupt, J. 1997. On the relationship between the Matriculation examination and University performance in South Africa—1980 to 1991. *South African Journal of Science*. 93, 382-387.

Morgan, J. N. & Messenger, R. C. 1973. THAID-a sequential analysis program for the analysis of nominal scale dependent variables. *Survey Research Centre, Institute for Social Research*, University of Michigan.

Nobel, J. & Sawyer, R. 1997. Alternative methods for validating admissions and course placement criteria. *The Association for Institutional Research for Management Research, Policy Analysis and Planning*. 63, 1-12.

Nunns, C. & Ortlepp, K. 1994. Exploring predictors of academic success in psychology 1 at Wits University as an important component of fair student selection. *South African Journal of Psychology*. 24, 201-207.

Pampel, F. C. 2000. Logistic regression: A primer. *Sage University Papers Series on Quantitative Applications in the Social Sciences, 07-132*. Thousand Oaks, CA:Sage.

Pascarella, E. T. & Terenzini, P. T. 2005. How College affects students: A third decade of research. San Francisco: Jossey-Bass.

Perreault, Jr. W. D. & Barksdale, Jr. H. C. 1980. A model-free approach for analysis of complex contingency data in survey research. *Journal of Marketing Research*. 17, 503-515.

Perreault, Jr. W. D. & Wagner, H. M. 1979. Applications and extensions of interaction analysis. Presented at Symposium on Interaction Analysis. ORSA/TIMS conference, New Orleans(May).

Power, C., Robertson, F. & Baker, M. 1987. Success in Higher Education. *Canberra: Australian Government Publishing Service.*

Raftery, A. E. 1995. Bayesian model selection in social research. In (P.V. Marsden, Ed.) *Sociological Methodology* (p. 111-163). London, Tavistock.

Ripley, B. D. 1996. Pattern recognition and neural networks. *Cambridge University Press.*

Robbins, S. B., Lauver, K., Le, H., Davis, D., Langley, R., & Carlstorm, A. 2004. Do psychosocial and study skill factors predict college outcomes? *Psychological Bulletin.* 130, 261-288.

Robinson, L. 1969. Relationship of student persistence in college to satisfaction with 'environmental' factor. *Journal of Educational Research*. 63, 6-10.

Roux, N. J. Le., Bothma, A. & Botha, H. L. 2004. Statistical properties of indicators of first-year performance at University. *The Journal of ORSSA.* 20, 161-178.

Spady, W. 1970. Dropouts from Higher Education: An interdisciplinary review and synthesis. *Interchange.* 1, 64-85.

Spady, W. 1971. Dropouts from higher education: Toward an empirical model. *Interchange.* 2, 38-62.

Stage, F. K. & Hossler, D. (2000). Where is the student?: Linking student behaviours, College choice and College persistence. In J. M. Braxton (Ed.), *Reworking the Student Departure Puzzle*, 170-195, Nashville: Vanderbilt University Press.

Stoker, D. J. 1985. Investigation into different entrance requirements for tertiary educational Institutions. *Report WS-32*, Human Sciences Research Council, Pretoria.

Stokes, M. E., Davis, C. S. & Koch, G. G. 2000. Categorical data analysis using the SAS system. *SAS Institute Inc.* Cary, NC, USA.

Tinto, V. 1975. Dropout from Higher Education: A theoretical synthesis of recent research. *Review of Educational Research*. 45, 89-125.

Wonnacott,T. H. & Wonnacott, R. J. 1981. Regression: A second course in statistics. *John Wiley & Sons*, USA.

Working group on retention and throughput. 2003. Success at the University of the Witwatersrand, *Report.*

Wright, R. E. 1995. Logistic regression. In L.G. Grimm, and P.R. Yarnold, Eds. *Reading and understanding multivariate statistics*, 217-244. American Psychological Association, Washington, DC.

Yeld, N. 2003. Academic literacy and numeracy profiles: An analysis of some results from the AARP and TELP tests of incoming students (2001/2002 entry years). In J. Withers and H. Griesel (Eds.), *Into Higher Education: Perspectives on Entry Thresholds and Enrolment Systems (p.21-52)*. Pretoria:SAUVCA-CTP

APPENDIX 1

DISTRIBUTION OF THE VARIABLES OF TRAINING AND VALIDATION DATA

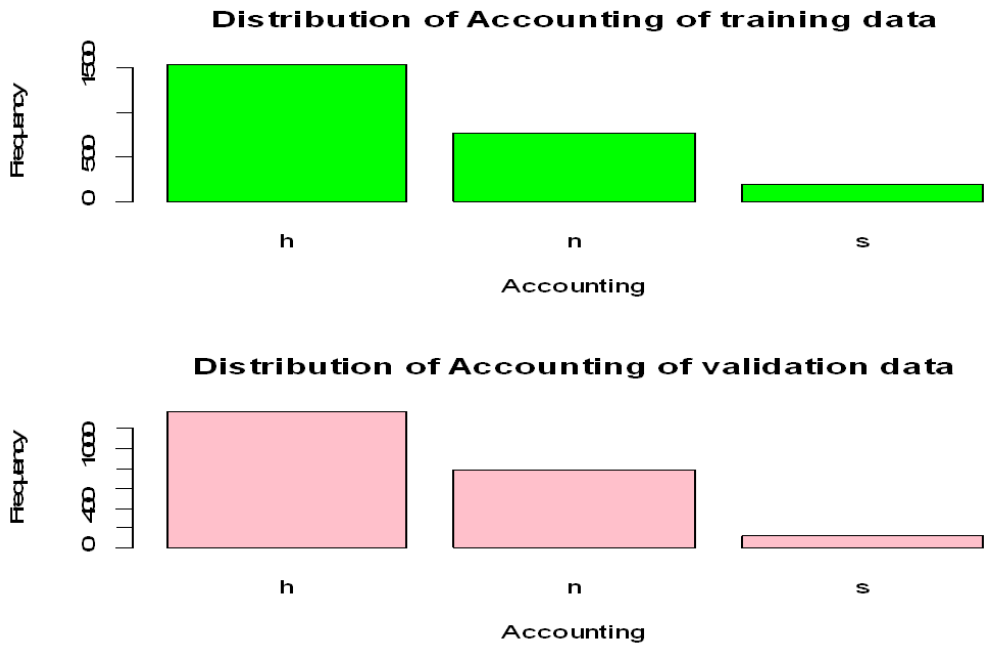

Figure A1.1 Distribution of the Accounting

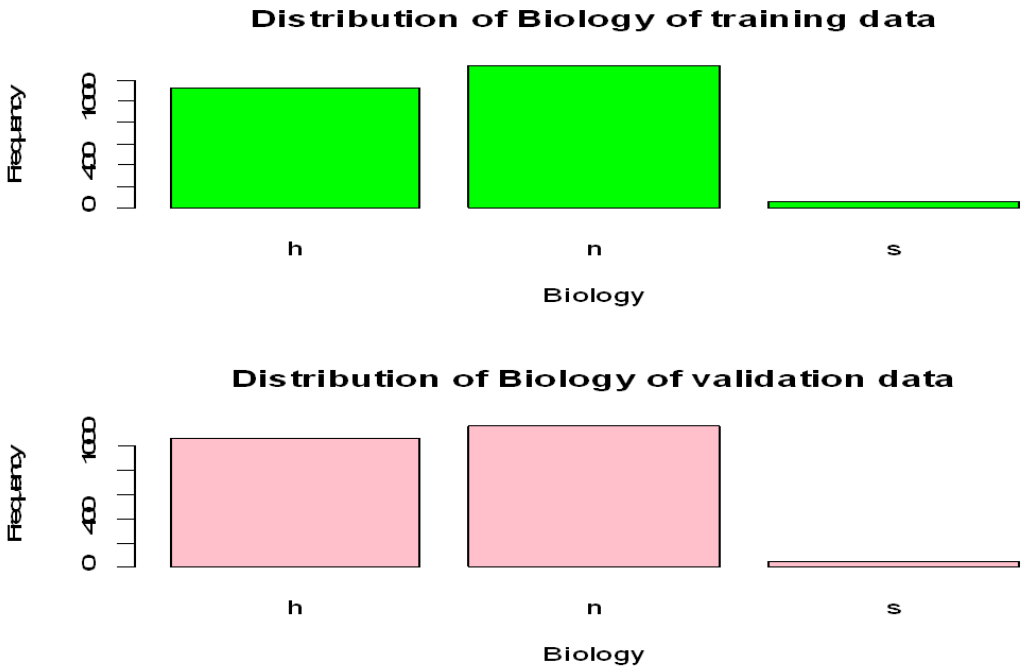

Figure A1.2 Distribution of the Biology

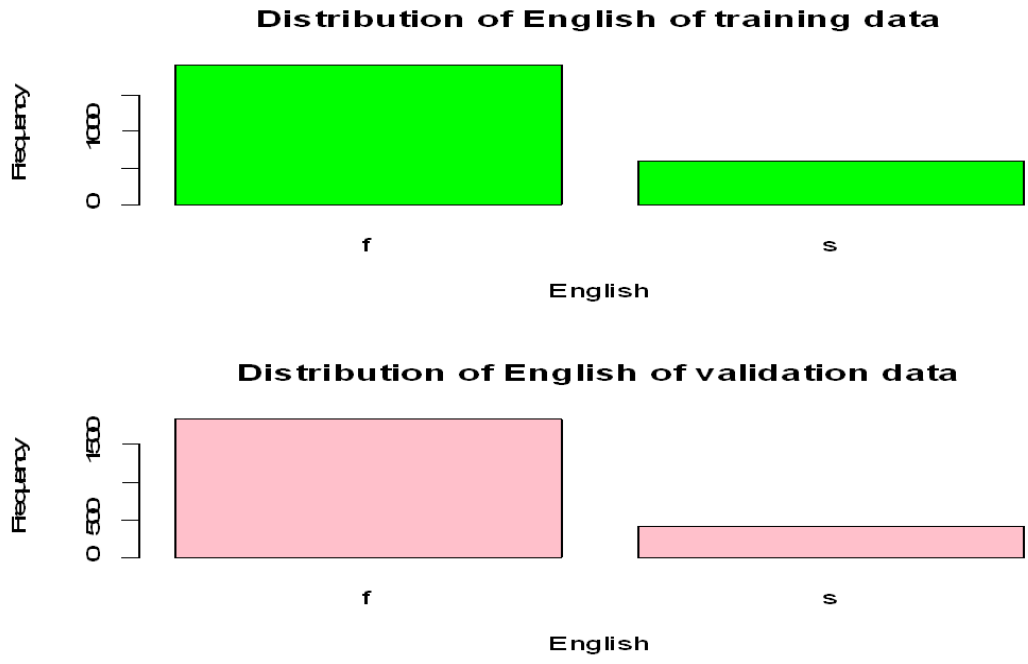

Figure A1.3 Distribution of the English

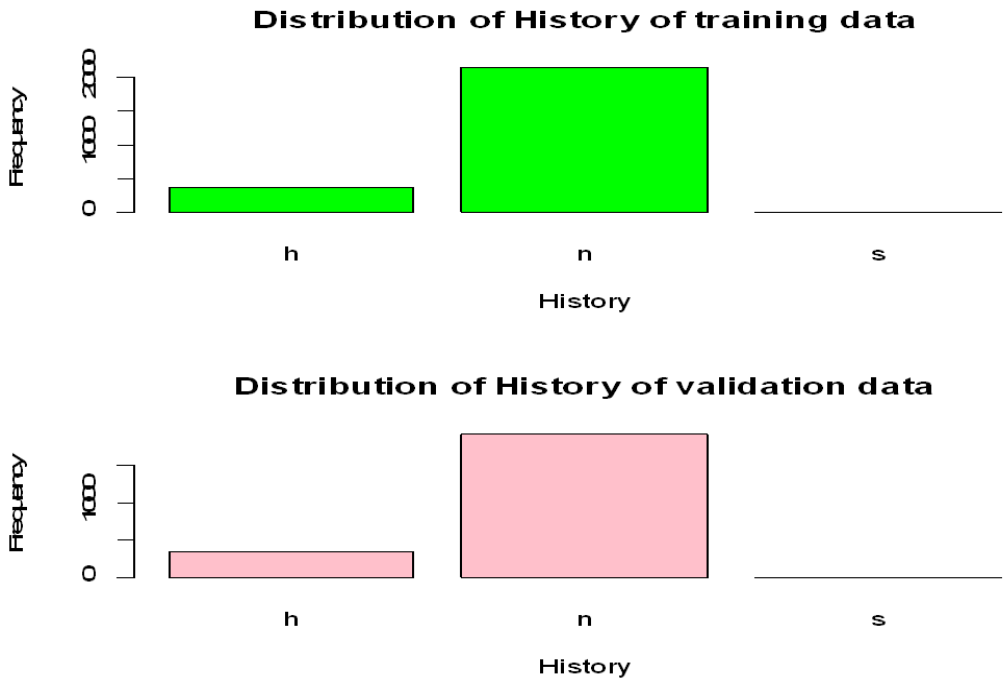

Figure A1.4 Distribution of the History

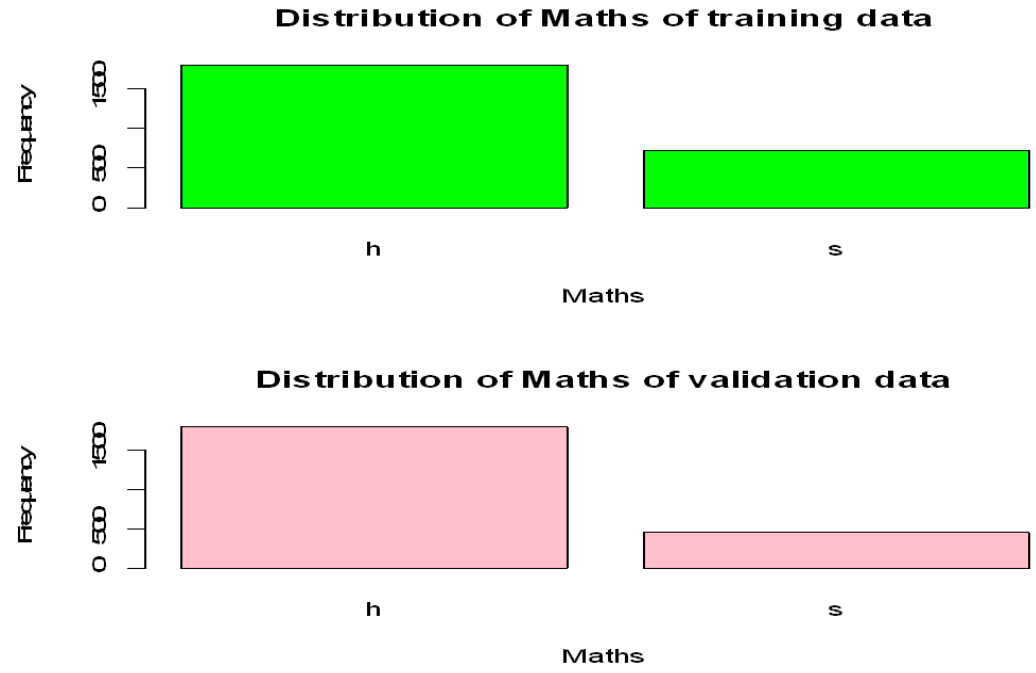

Figure A1.5 Distribution of the Mathematics

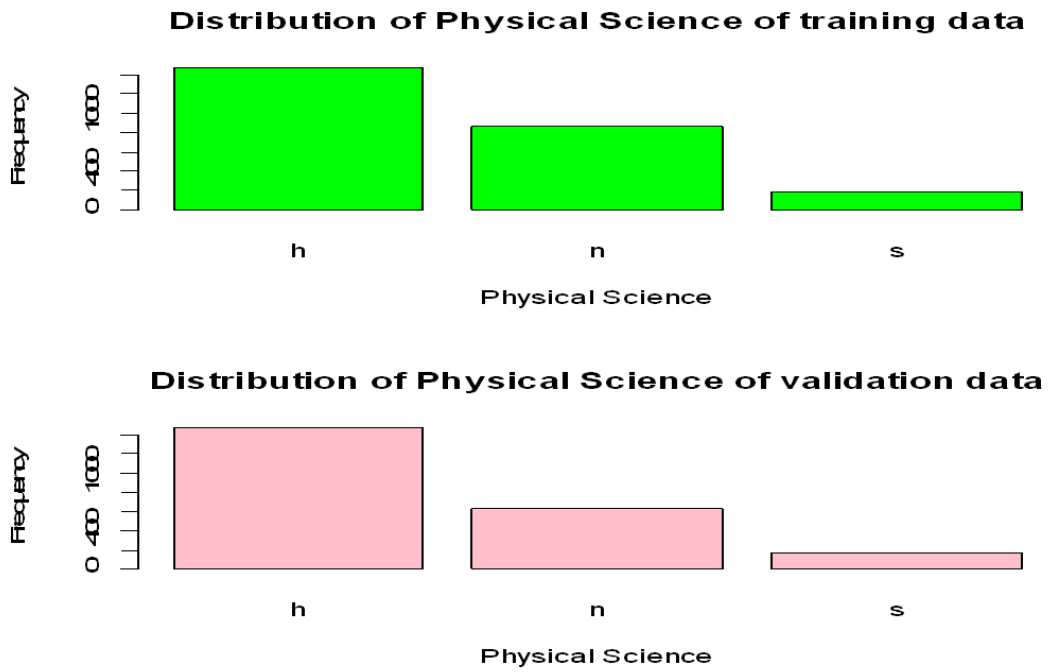

Figure A1.6 Distribution of the Physical Science

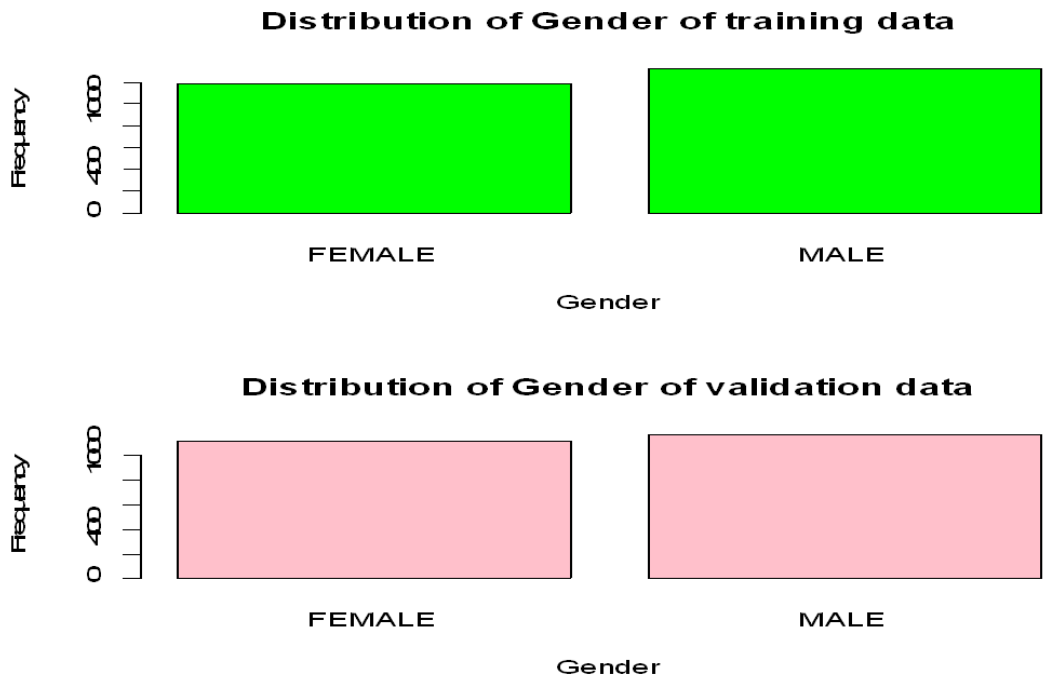

Figure A1.7 Distribution of the Gender

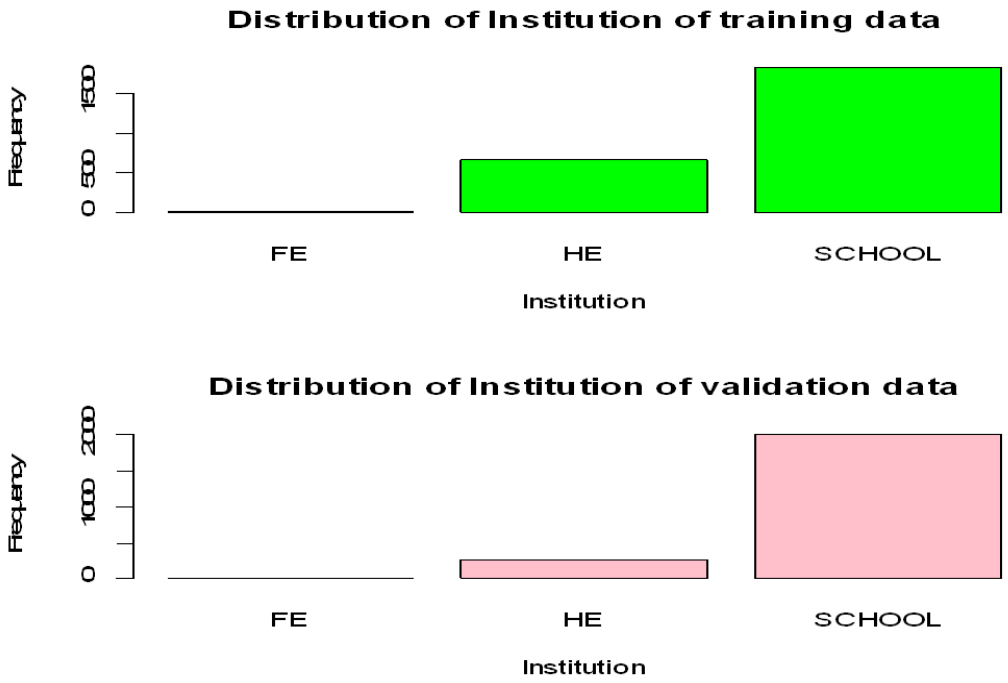

Figure A1.8 Distribution of the Previous Institution Type

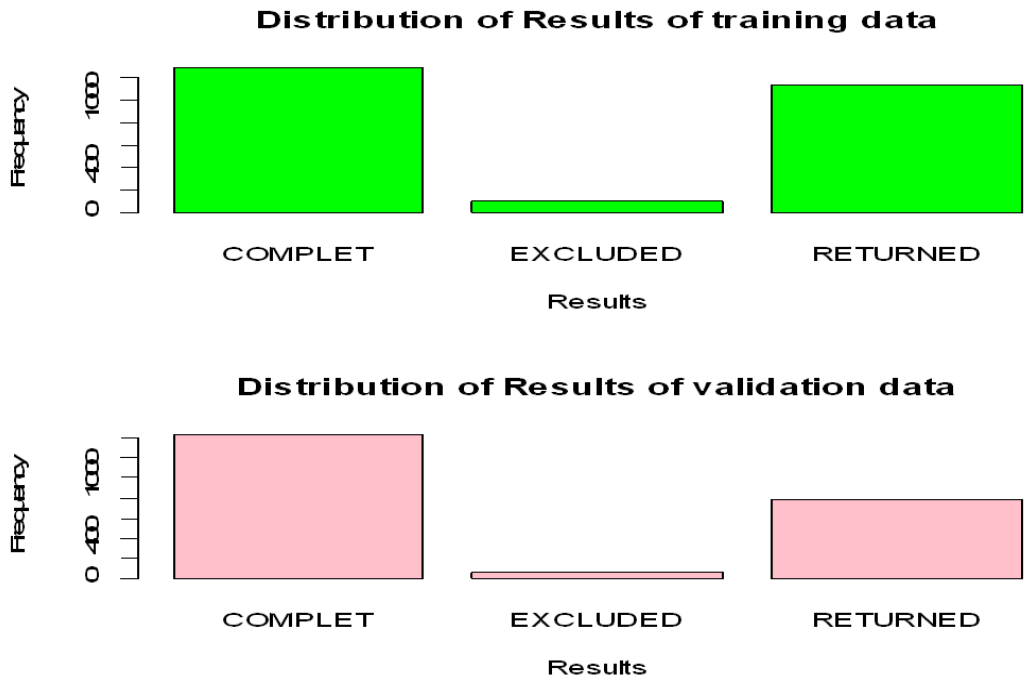

Figure A1.9 Distribution of the First Year Performance

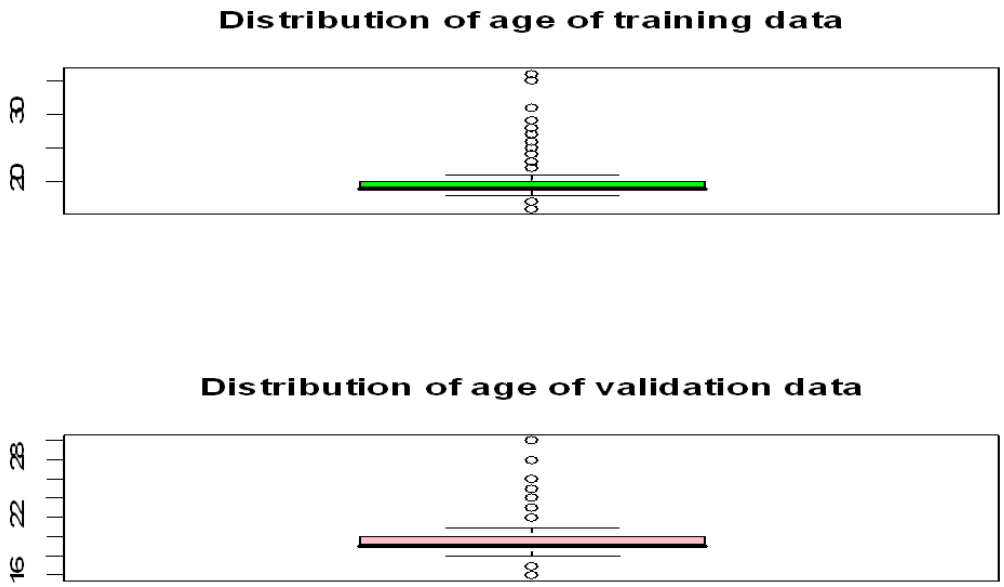

Figure A1.10 Distribution of the Age

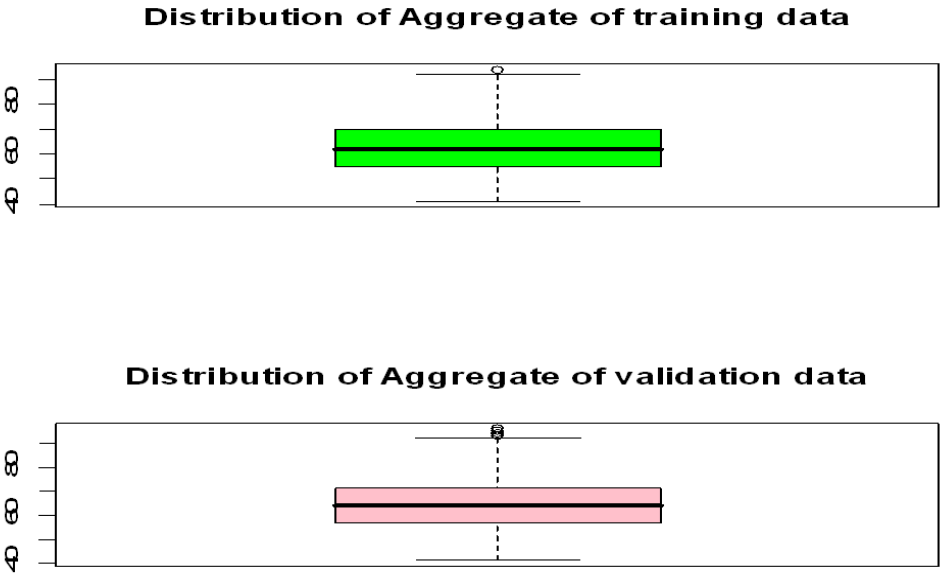

Figure A1.11 Distribution of the Aggregate

APPENDIX 2

STUDENTS' PERFORMANCE BY AGE AND AGGREGATE OF TRAINING AND VALIDATION DATA

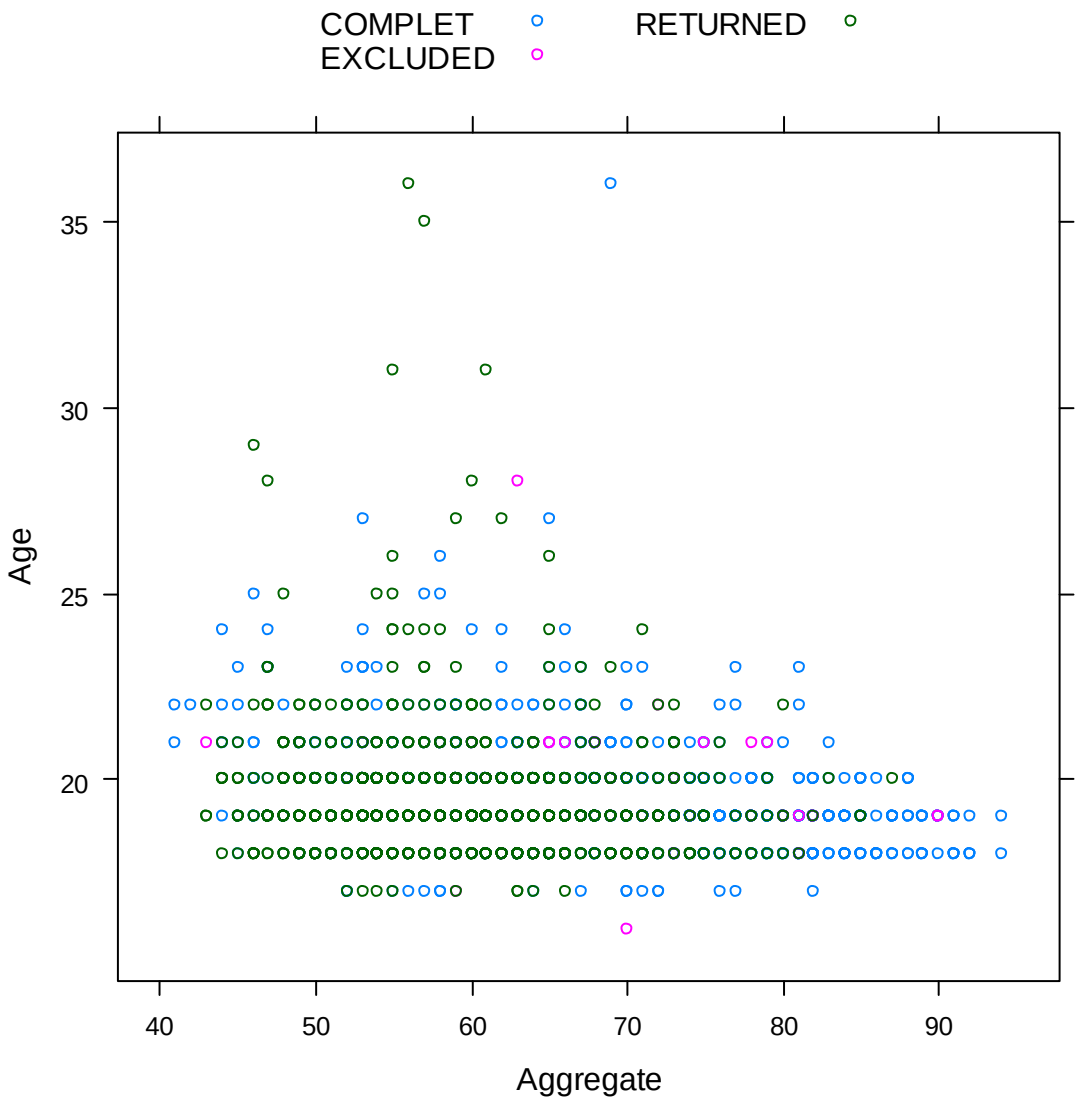

Figure A2.1: First year Performance by Age and Aggregate of training data

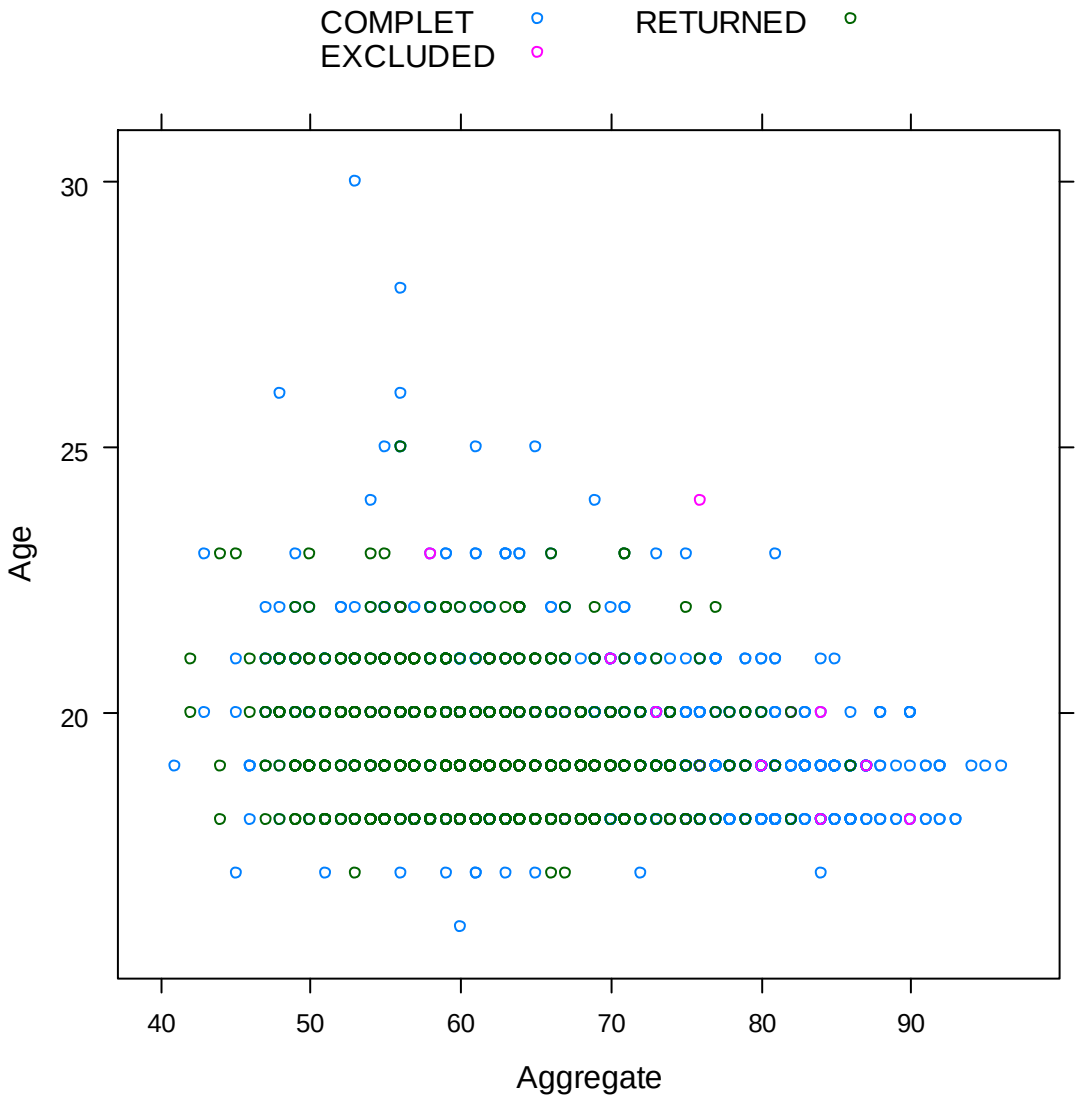

Figure A2.2: First year Performance by Age and Aggregate of validation data