LIBRARY
UNIVERSITY OF MORATUWA, SRI LANKA
MORATUWA

(18)

28/2011/37/20

An Investigation on Factors Affecting the Further Education Desires of Youth in Sri Lanka
&
A Decision Support Tool for Data Mining

By
D.R.T. Jayasundara
06/8105

University of Moratuwa, Sri Lanka.
Electronic Theses & Dissertations
www.lib.mrt.ac.lk

This dissertation is submitted in partial fulfillment of the requirements for the M.Sc. Degree in Operational Research.

Department of Mathematics
University of Moratuwa
Sri Lanka

October 2009

51 "09"
519.8(043)

96437

TH

University of Moratuwa
96437

96437

UNIVERSITY OF MORATUWA
LIBRARY
SRI LANKA

Declaration

I officially state that this dissertation titled "An Investigation on Factors Affecting the Further Education Desires of Youth in Sri Lanka and a Decision Support Tool for Data Mining" is entirely my own work except where explicitly mentioned. It has not been submitted nor being currently submitted for any other degree.

Candidate : Ms. D. R. T. Jayasundara

Signature:.. *UOM Verified Signature*

Date :..01/03/2010

Supervisor : Dr. M.D.T. Attygalle
Senior Lecturer
Department of Statistics,
Faculty of Science,
University of Colombo.

Signature:.. *UOM Verified Signature*

Date :........01/03/2010

University of Moratuwa, Sri Lanka.
Electronic Theses & Dissertations
www.lib.mrt.ac.lk

Supervisor : Mr. M. Firdhous
Senior Lecturer
Department of Information Technology,
Faculty of Information Technology,
University of Moratuwa.

Signature:. *UOM Verified Signature*

Date :..01/03/2010.

Coordinator Mr. T.M.J.A. Cooray
Senior Lecturer
Department of Mathematics,
Faculty of Engineering,
University of Moratuwa.

Signature:............ *UOM Verified Signature*

Date :...01/03/2010

To My Beloved Husband &Son Geethaka

With

Heartfelt Love and Admiration.

University of Moratuwa, Sri Lanka.
Electronic Theses & Dissertations
www.lib.mrt.ac.lk

Abstract

Education plays an important role in the development of a country. Sri Lanka has given priority to widen access to education and has taken measures but whether everyone gets the equal opportunity to education is still open to question. This has been considered in an island wide National Youth Survey, conducted in the year 2000 which has been a joint undertaking involving United Nations Development Program (UNDP) and six Sri Lankan and German Institutions. The main aim of the survey has been to collect up-to-date and reliable information about opinions, values, perceptions, concerns, grievances and aspirations of the young generation in Sri Lanka. Further it has aimed at the identification and better understanding of the main commitment and ideas to solve them. It has been intended to provide a scientific database to assist policy makers and development organizations. The data set collected in this survey has data on specific segments related to youth in the age spectrum 15-29 years. This set of data is subjected to a social research. In this study, statistical evidence is sought to follow the proceedings of data mining.

A statistical modeling approach has been used in the analysis. The type of Further Educational Desire of a person has been found to be mainly related with the Type of Current Activity in terms of current employment status and so on, Educational Level, Province, Gender, Social Class and Age Group. Moreover, sufficient statistical evidence has been available to say that even, the Financial Situation in Past and the Major Problems Occurred in Education have an effect on developing their further educational desires. The importance of these findings is that they can be useful to understand the facilities and opportunities to be provided for students to assist in their education and to achieve their ultimate educational and career goals. Besides that, the information could be useful to assist in educational reforms and policy making.

A decision support tool has been developed using the rule based method in data mining using the information furnished by the inferential statistical analysis. The social characteristics "Type of Current Activity", "Educational Level", "Province", "Gender", "Social Class" and "Age Group" were found to be directly influential in predicting the type of further educational desires of youth and they were used in developing the decision support tool. It predicts what type of educational desires can be there with the individuals in a selected sample of youth. These predictions can be observed by any individual education provider or an organization before launching their educational centers targeting the selected population.

University of Moratuwa, Sri Lanka.
Electronic Theses & Dissertations
www.lib.mrt.ac.lk

Acknowledgement

I wish to express my appreciation to the course coordinator, Mr. T.M.J.A. Cooray, the Head of the Department of Mathematics, Faculty of Engineering, University of Moratuwa, who has given his best assistance and advice towards the success of this project.

Words cannot express my gratitude to my supervisors Dr. Dilhari Attygalle and Mr. M. Firdhous, without whose expert advice, guidance and help this study would not have been easily get succeeded.

I specially thank Ms. Geethanjalee Henadheera for making me to do an important analysis on Youth Education in Sri Lanka by providing the data set.

I thankfully mention the support given by the staff members of Social Policy Analysis & Research Centre, Faculty of Arts, University of Colombo by giving valuable information on the data set and the questionnaire.

University of Moratuwa, Sri Lanka.
Electronic Theses & Dissertations
www.lib.mrt.ac.lk

My Sincere gratitude goes to Prof. Roshini Sooriyarachchi, Ms. A. Karunaratne and Mr. R. A. B. Abeygunawardane, senior lecturers in the Department of Statistics, University of Colombo for guiding me to complete this project work punctually, overcoming all obstacles.

Last but not least, my heartfelt thanks go to all my friends who were always with me in-need and helped me in many ways.

Table of Contents

Abstract.........i
Acknowledgement.........ii
List of Tables.........v
List of Figures.........vi
Figure 1.4.1: Link between Data Miners, Statisticians & Computer Scientists 6 vi
1 Chapter 1.........1
Introduction 1
1..1 Background of the Study.........2
1..2 About the Survey.........3
1..3 Sample Design.........3
1..4 Objectives of the Study 4
1..5 Significance of the Study 5
1..6 Link between Data Mining, Statistics, Computer Science.........6
Figure 1.4.1: Link between Data Miners, Statisticians & Computer Scientists.........6
1..7 Organization of the Dissertation 9
2 Chapter 2.........10
Review of the Social Research 10
3 Chapter 3.........15
Literature review of Data mining 15
3..1 Data Mining Concept 16
3..2 Data Mining Techniques 17
3..3 Data Mining Tools 19
3..4 Data Mining Applications 20
4 Chapter 4.........24
Theory and Methodology 24
4..1 Methodology Used in Preliminary Analysis 25
4..2 Introduction to the Polytomous Regression Model.........25
4..3 Model Building Strategies for Polytomous Logistic Regression 26
4..4 Generalized Logit Model 27
4..5 Model Selection Procedure 28
4..6 Goodness of Fit and Diagnostics for Polytomous Logistic Regression 28
4..7 Goodness of Fit of the Logistic Model.........29
4..8 Data Mining (DM) 30
5 Chapter 5.........35

University of Moratuwa, Sri Lanka.
Electronic Theses & Dissertations
www.lib.mrt.ac.lk

Preliminary Analysis 35
5..1 Descriptive Analysis 36
5..2 Univariate Analysis 50
6 Chapter 6....................52
Advanced statistical analysis 52
6..1 Fitting a Statistical Model 53
6..2 About Data 53
6..3 Fitting the Best Fitted Generalized Logit Model 54
6..4 Checking the Adequacy of the Best 2-way Interaction Model 68
7 Chapter 7....................70
Implementation of Data Mining Techniques....................70
7..1 Introduction to Data 71
7..2 Construction of the Decision Tree 71
7..3 Deployment 76
7..4 Evaluation....................77
8 Chapter 8....................83
General Discussion And Conclusion....................83
8..1 The Availability of Data....................84
8..2 The Quality of Data....................84
8..3 Statistical Techniques....................85
8..4 Limitations to Further Work 85
8..5 Recommendations for Future Work....................85
8..6 Conclusion....................86

Bibliography

Appendix A - SAS output of the statistical model with parameter estimates

Appendix B - SAS codes

Appendix C – Reference to decision rules

Appendix D – Survey questionnaire

University of Moratuwa, Sri Lanka.
Electronic Theses & Dissertations
www.lib.mrt.ac.lk

List of Tables

Table 1.1.2.1: Comparison of Sample and Population figures by Age 4
Table 1.1.2.2: Comparison of Sample and Population figures by Gender 4
Table 4.1.1: Contingency table of variable A (with i levels) and variable B (with j levels) 26
Table 4.1.2: Cross tabulation for two variables 27
Table 4.2.1: Confusion Matrix for a binary classifier 36
Table 4.2.2: Confusion Matrix for a multi classifier 37
Table 5.1.1: The frequency table for Age Group 40
Table 5.1.2: The frequency table for Gender 41
Table 5.1.3: The frequency table for Educational Level 41
Table 5.1.4: The frequency table for Ethnicity 42
Table 5.1.5: The frequency table for Province 43
Table 5.1.6: The frequency table for Sector 44
Table 5.1.7: The frequency table for Social Class (Self-Identification) 45
Table 5.1.8: The frequency table for Present Financial Status 46
Table 5.1.9: The frequency table for Financial Situation in Past 47
Table 5.1.10: The frequency table for whether schools provide good education 48
Table 5.1.11: The frequency table for Major Problems with Education 49
Table 5.1.12: The frequency table for Access to Educational Facilities 50
Table 5.1.13: The frequency table for Type of Activity 51
Table 5.2.1: Results of assessing the significance of variables at 20% sig. level. 54
Table 6.2.1: Reference to Explanatory Variables (significant at 20% level) 57
Table 6.3.1: Adding the first most significant variable to the model 58
Table 6.3.2: Adding the second most significant variable to the model 59
Table 6.3.3: Adding the third most significant variable to the model 60
Table 6.3.4: Adding the fourth most significant variable to the model 61
Table 6.3.5: Adding the fifth most significant variable to the model 61
Table 6.3.6: Adding the sixth most significant variable to the model 62
Table 6.3.7: Fitting the first most significant 2-way interaction terms to the best main effect model 64
Table 6.3.8: Fitting the second most significant 2-way interaction term to the model 66
Table 6.3.9: Fitting the next most significant 2-way interaction term to the Model 67
Table 6.4.1: Testing the Goodness of Fit 68
Table 6.4.2: Classification table of Model 9 68
Table 7.1.1: Selected attributes for Data Mining 71
Table 7.3.1:Used Attributes for data mining model 76
Table 7.4.1: Confusion Matrix of three category response variable 77
Table 7.4.2: Confusion Matrix adjusted for No Desire 77
Table 7.4.3: Confusion Matrix adjusted for University/Equivalent Higher Institution Education 78
Table 7.4.4: Confusion Matrix adjusted for Technical/Vocational Education 78
Table 7.4.5: Predicted and Observed Frequency Table for the Four Subsets of the Test Data Set 79
Table 7.4.6: Confusion Matrix Adjusted for No Desire for Four Subsets of Test Data 79

University of Moratuwa, Sri Lanka.
Electronic Theses & Dissertations
www.lib.mrt.ac.lk

Table 7.4.7: Confusion Matrix Adjusted for University/Equivalent Higher Institution Education for Four Subsets of Test Data 80
Table 7.4.8: Confusion Matrix Adjusted for Technical/Vocational Education for Four Subsets of Test Data 80
Table 7.4.9: Measures Obtained from the Four Subsets of Test Data 81

List of Figures

Figure 1.4.1: Link between Data Miners, Statisticians & Computer Scientists 6
Figure 1.4.2.1:Data Mining (KDD) Process 7
Figure 1.4.2.2:Data Preparation Process 7
Figure 1.4.2.3:Mining and Modeling Process 8
Figure 1.4.2.4:Consolidation and Application 8
Figure 4.3.1: Available Data Mining Techniques 34
Figure 5.1.1: Distribution of Respondents among Different Further Education Desires 39
Figure 5.1.2: Distribution of Respondents among Different Further Education Desires by Age 40
Figure 5.1.3: Distribution of Respondents among Different Further Education Desires by Gender 41
Figure 5.1.4: Distribution of Respondents among Different Further Education Desires by Educational Level 42
Figure 5.1.5: Distribution of Respondents among Different Further Education Desires by Ethnicity 43
Figure 5.1.6: Distribution of Respondents among Different Further Education Desires by Province 44
Figure 5.1.7: Distribution of Respondents among Different Further Education Desires by Sector 45
Figure 5.1.8: Distribution of Respondents among Different Further Education Desires by Social Class (Self-Identification) 46
Figure 5.1.9: Distribution of Respondents among Different Further Education Desires 47
Figure 5.1.10:Distribution of Respondents among Different Further Education Desires by Financial Situation in Past 48
Figure 5.1.11:Distribution of Respondents among Different Further Education Desires by Whether schools provide good education 49
Figure 5.1.12: Distribution of Respondents among Different Further Education Desires by Major Problems with Education 50
Figure 5.1.13:Distribution of Respondents among Different Further Education Desires by Access to Educational Facilities 51
Figure 5.1.14:Distribution of Respondents among Different Further Education Desires by Type of Activity 52
Figure 5.1.1: Distribution of Respondents among Different Further Education Desires
Figure 7.2.1: Sample Decision Tree 73
Figure 7.2.2: Decision Tree for Training Data in Table Format Generating rules from a Decision Tree 74
Figure 7.3.1: User friendly implementation of the model 76
Figure 7.4.1: ROC plot of the DM model 82

University of Moratuwa, Sri Lanka.
Electronic Theses & Dissertations
www.lib.mrt.ac.lk

University of Moratuwa, Sri Lanka.
Electronic Theses & Dissertations
www.lib.mrt.ac.lk