

# **PREDICTIVE MODELING ON TELEKOM MALAYSIA DIRECT EXCHANGE LINE GROWTH**

**A thesis submitted to the Graduate Study Centre in partial**

**fulfillment of the requirements for the degree**

**Master of Science (Intelligent System)**

**Universiti Utara Malaysia**

**By**

**Roznim Binti Mohamad Rasli**

**© Roznim Binti Mohamad Rasli, July 2005. All rights reserved**



**JABATAN HAL EHWAL AKADEMIK**  
**(Department of Academic Affairs)**  
**Universiti Utara Malaysia**

**PERAKUAN KERJA KERTAS PROJEK**  
**(Certificate of Project Paper)**

Saya, yang bertandatangan, memperakukan bahawa  
*(I, the undersigned, certify that)*

**ROZNIM BINTI MOHAMAD RASLI**

calon untuk Ijazah  
*(candidate for the degree of)* **MSc. (Int. Sys)**

telah mengemukakan kertas projek yang bertajuk  
*(has presented his/ her project paper of the following title)*

**PREDICTIVE MODELING ON TELEKOM MALAYSIA**  
**DIRECT EXCHANGE LINE GROWTH**

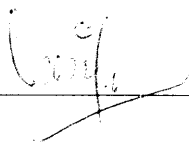
seperti yang tercatat di muka surat tajuk dan kulit kertas projek  
*(as it appears on the title page and front cover of project paper)*

bahawa kertas projek tersebut boleh diterima dari segi bentuk serta kandungan dan meliputi bidang ilmu dengan memuaskan.  
*(that the project paper acceptable in form and content, and that a satisfactory knowledge of the filed is covered by the project paper).*

Nama Penyelia Utama  
*(Name of Main Supervisor):* **ASSOC. PROF. DR. NORITA MD. NORWAWI**

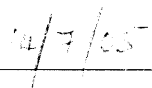
Tandatangan  
*(Signature)*

:

  
\_\_\_\_\_

Tarikh  
*(Date)*

:

  
\_\_\_\_\_

## **PERMISSION TO USE**

In presenting this thesis in partial fulfillment of the requirements for a postgraduate degree from Universiti Utara Malaysia, I agree that the University Library may make it freely available for inspection. I further agree that permission for copying of this thesis in any manner, in whole or in part, for scholarly purpose may be granted by my supervisor(s) or, in their absence by the Director of Graduate Study Centre. It is understood that any copying or publication or use of this thesis or parts thereof for financial gain shall not be allowed without my written permission. It is also understood that due recognition shall be given to me and to Universiti Utara Malaysia for any scholarly use which may be made of any material from my thesis.

Requests for permission to copy or to make other use of materials in this thesis, in whole or in part, should be addressed to

**Director of Graduate Study Centre**

**Universiti Utara Malaysia**

**06010 UUM Sintok**

**Kedah Darul Aman.**

# ABSTRAK

(Bahasa Melayu)

Perlombongan data merujuk kepada pengekstrakan maklumat-maklumat peramalan yang tersembunyi di dalam pangkalan data dan telah menjadi salah satu teknologi penting yang berpotensi untuk membantu syarikat memfokuskan kepada maklumat-maklumat penting di dalam gudang data mereka. Model peramalan menjadikan peramalan tentang nilai-nilai data berdasarkan kepada maklumat-maklumat atau hasil-hasil telah diperolehi daripada data-data terdahulu di mana kemungkinan hasil yang terbaik adalah berdasarkan kepada perolehan data sebelumnya. Telekom Malaysia (TM) adalah merupakan pelopor kepada era telekomunikasi di Malaysia yang telah membekalkan kemudahan-kemudahan komunikasi dan juga bertindak sebagai tulang belakang digital. Talian Ibusawat Terus (*Direct Exchange Line* – DEL) adalah salah satu kemudahan telefoni utama di TM yang mengendalikan sejumlah besar dan pelbagai jenis data di dalam operasi hariannya. Oleh itu, amat sukar untuk menyerlahkan struktur-struktur pengetahuan yang dapat membantu dalam pembuatan keputusan terutamanya dalam keadaan kebarangkalian yang begitu terhad. Matlamat utama tesis ini adalah untuk mengenalpasti teknik perlombongan data yang paling sesuai di antara regresi logistik (*logistic regression*), pohon keputusan (*decision tree*) dan rangkaian neural (*neural networks*) untuk meramal peningkatan dan perkembangan DEL berdasarkan kepada lima atribut-attribut fizikal penting yang terdiri daripada ibusawat, pelanggan, pemasangan baru, pemotongan, dan ketersediaan kabel atau port (ECP) yang menyumbang kepada 672 rekod yang mensasar kepada samada berlaku peningkatan atau penurunan dalam perkembangan DEL di TM khususnya di Pulau Pinang. Keputusan daripada peramalan ini amat penting untuk mendapatkan pemahaman yang lebih mendalam tentang masa hadapan pasaran berdasarkan kepada situasi semasa dan terdahulu.

## **ABSTRACT**

**(English)**

Data mining (DM) is the extraction of hidden predictive information from large databases that has becoming a powerful new technology with great potential to help companies focus on the most important information in their data warehouses. A predictive model makes a prediction about values of data using known results found from historical data where the best possible outcome based on the previous data is derived. Telekom Malaysia (TM) is Malaysia's premier communications provider that provides the digital backbone and communication facilities. Direct Exchange Line (DEL) is one of its core telephony services that handle massive volume and variety of data in its daily operations. Therefore, it is hard to reveal knowledge structures that can guide decisions in conditions of limited certainty. The main objective of this study is to identify the most appropriate DM technique between logistic regression, decision tree and neural networks for predicting DEL growth based on five physical attributes namely exchanges, subscribers, new installation, cutting, and availability of cable or ports (ECP) that constitute of 672 instances leading to a target (either increase or decrease). The result of this study is important in assisting the prediction of DEL growth in TM specifically in Penang, thus leading on better understanding on the future of the market based on the current and previous situation.

## ACKNOWLEDGEMENTS

I would like to gratefully acknowledge the contributions of several people who have helped me in completing this thesis. Without them, I am nothing here.

First and foremost, praise be to Allah SWT, whose blessing and guidance have helped me in completing this thesis. Peace be upon our Prophet Muhammad S.A.W, who has given light to mankind. My sincere appreciations to my beloved parents (Mr. Haji Mohamad Rasli Bin Haji Hashim and Mrs. Aminah Binti Hasan) and family for their patience, prayers and understanding over the entire period of my study.

Second, I would like to convey my grateful thanks to my talented supervisor, Associate Professor Dr. Norita Md. Norwawi that has given a full support and has contributed immensely towards the completion of this thesis. She has spent all her time patiently and painstakingly giving advices and valuable information, correcting errors in training process, editing the text, and proofreading the thesis just to ensure the best effort has been given in the completion and achievement of this thesis. Needless to say, I could not have completed writing this thesis if its not been for her admirable diligence and resourcefulness.

Third, I would like to express my gratitude to TM staffs especially to Mr. Haji Mazlan for his outstanding ideas in considering Telekom Malaysia (TM) especially its core service,

Direct Exchange Line (DEL), as my main subject. Without him, there would be no thesis entitled “Predictive Modeling on Telekom Malaysia Direct Exchange Line Growth”.

Fourth, I would like to express my special thanks to Mr. Haji Mohamad Rasli Bin Haji Hashim, that have largely been responsible in making the data acquisition process runs smoothly. He also has spent all his time patiently and painstakingly giving advices and valuable information just to ensure that I have all the stuffs that I need. Thanks dad for all your help, your full supports and your trust. I don't think I managed to complete this thesis without his support and encouragement.

Equally deserving of this recognition are brilliant Intelligent System lecturers who have supported me towards the completion of this thesis especially Mr. Wan Hussein Wan Ishak and Associate Professor Azizi Zakaria.

Last but not least, to all my beloved friends and all the individuals that have involve directly and indirectly in the completion of this thesis. All their support and encouragement is very much appreciated. Thanks to all.

Roznim Binti Mohamad Rasli

July, 2005

# CONTENTS

PERMISSION TO USE	i
ABSTRAK (BAHASA MELAYU)	ii
ABSTRACT (ENGLISH)	iii
ACKNOWLEDGEMENTS	iv
CONTENTS	vi
LIST OF TABLES	ix
LIST OF FIGURES	x
LIST OF ABBREVIATIONS	xi
<b>CHAPTER ONE : INTRODUCTION</b>	
1.1 Problem Statement	3
1.2 Objective	3
1.3 Scope	4
1.4 Significance of Study	5
1.5 Organization of the Report	6
1.6 Summary	7
<b>CHAPTER TWO : LITERATURE REVIEW</b>	
2.1 Data Mining (DM)	8
2.2 Predictive Modeling	11
2.3 Techniques within Data Mining (DM)	12
2.4 Summary	15



## **CHAPTER THREE : METHODOLOGY**

3.1 Methodology	16
3.1.1 Phase 1 : Business Understanding	17
3.1.2 Phase 2 : Data Understanding	20
3.1.3 Phase 3 : Data Preparation	24
3.1.4 Phase 4 : Modeling	25
3.1.5 Evaluation	26
3.2 Summary	26

## **CHAPTER FOUR : MODELING AND EVALUATION**

4.1 Modeling	28
4.1.1 Selection of modeling technique	28
4.1.2 Generating test design	29
4.1.3 Building the model	35
4.2 Evaluation	36
4.2.1 Model (or knowledge) representation	37
4.2.2 Estimation criterion	37
4.2.3 Search method	58
4.3 Summary	58

## **CHAPTER FIVE : DISCUSSION**

5.1 Results	59
5.1.1 Comparison of Results	60

5.2 Summary	65
<b>CHAPTER SIX : CONCLUSION</b>	
6.1 Project Review	66
6.2 Contribution	68
<b>REFERENCES</b>	70
<b>APPENDIX A : Raw Data of DEL</b>	72

## LIST OF TABLES

<b>Table No.</b>	<b>Name of Table</b>	<b>Page No.</b>
Table 3.1	Proposed attributes of DEL data set	21
Table 3.2	The descriptions of 7 physical attributes and a target	22
Table 3.3	The conceptual model of DEL data set	23
Table 4.1	Selection of modeling techniques / classifiers	29
Table 4.2	Data partition summarization of DEL data set	30
Table 4.3	Determination results of HU	32
Table 4.4	Determination results of LRate	34
Table 4.5	Determination results of MR	35
Table 4.6	Summarization results for the most suitable parameters in NN (MLP)	35
Table 4.7	The interpretation of resultant models with different data partition	36
Table 5.1	The estimation criterion for the DEL increment based on the potential exchanges	63
Table 5.2	The estimation criterion of DEL data set	64

## LIST OF FIGURES

<b>Figure No.</b>	<b>Name of Figure</b>	<b>Page No.</b>
Figure 3.1	Phases in DM	17
Figure 3.2	Summarization of Possibility Study	19
Figure 4.1	The resultant of cumulative % response lift chart	40
Figure 4.2	The resultant of diagnostic chart	41
Figure 4.3	The resultant of confusion matrix	42
Figure 4.4	The resultant of assessment table	44
Figure 4.5	The resultant of assessment plot	44
Figure 4.6	The resultant of attribute selection measure	45
Figure 4.7	The tree diagram	46
Figure 4.8	The resultant of cumulative % response lift chart	47
Figure 4.9	The resultant of diagnostic chart	48
Figure 4.10	The resultant of confusion matrix	49
Figure 4.11	The resultant of cumulative % response lift chart	53
Figure 4.12	The resultant of diagnostic chart	54
Figure 4.13	The resultant of confusion matrix	55
Figure 4.14	The network architecture	57
Figure 5.1	The resultant of cumulative % response	61
Figure 5.2	The resultant of cumulative expected profit values	62

## LIST OF ABBREVIATIONS

<b>Acronym</b>	<b>Meaning</b>
AI	Artificial Intelligence
AIM	Air Itam
BF	Batu Ferringhi
BI	Balik Pulau
BM	Bukit Mertajam
BMU	Batu Maung
BP	Backpropogation
BTH	Bukit Tengah
BW	Butterworth
BYB	Bayan Baru
CRISP-DM	Cross-Industry Standard Process for Data Mining
DM	Data Mining
DT	Decision Tree
ECP	Availability of Cables or Ports
EDA	Exploratory Data Analysis
GEEs	Generalized Estimating Equations
GLR	Gelugor
GPA	Grade Point Averages
GPU	Guar Perahu
HU	Hidden Units

JTG	Jelutong
KBS	Kepala Batas
KMR	Komtar
LR	Logistic Regression
LRate	Learning Rate
MGB	Machang Bubok
MLP	Multi-Layer Perceptron
MR	Momentum Rate
NN	Neural Networks
NT	Nibong Tebal
PG	Penang
SGD	Sungai Dua
SI	Sungai Bakap
SJA	Seberang Jaya
SZ	Simpang Ampat
TAT	Telok Air Tawar
Telco	Telecom Company
TGR	Tasek Gelugor
TJB	Tanjung Bungah
TKR	Telok Kumbar
TM	Telekom Malaysia

## **CHAPTER ONE**

### **INTRODUCTION**

The fierce nature of today's competitive landscape has forced organizations to operate more efficiently, not just on a day-to-day basis but also in planning for the future, thus competitive advantage requires more than estimates and educated guesses. To succeed and grow, organizations need an accurate picture of the future and the ability to reliably measure the impact of economic and marketplace factors. Strategic business planning requires the ability to model and simulate any business process and the factors that has an impact on those processes, no matter how complex it is.

Telekom Malaysia, hereafter called TM is Malaysia's premier communications provider, providing the digital backbone and communication facilities so essential in propelling Malaysia forward towards a first world environment. As Malaysia's only full service telecom company (telco), TM offers a comprehensive range of communication solutions in Direct Exchange Line (DEL), voice telephony, mobile, high speed broadband and internet services, multimedia applications, data services, broadcasting, audio and video conferencing and consultancy, thus TM already collects and refines massive quantities of data in its daily operations. Therefore, it is hard to reveal knowledge structures that can guide decisions in conditions of limited certainty.

The contents of  
the thesis is for  
internal user  
only



## REFERENCES

- Apte, C., Liu, B., Pednault, P.D., and Myth, P. (2002). Business Applications of Data Mining. Retrieved 9 April 2005 from [http://www.research.ibm.com/dar/papers/pdf/business\\_applications\\_of\\_dm.pdf](http://www.research.ibm.com/dar/papers/pdf/business_applications_of_dm.pdf)
- Berson, A., Smith, S., and Thearling, K. (2000). An Overview of Data Mining at Dun and BradStreet. Retrieved 9 April 2005 from <http://www.thearling.com/text/wp9501/wp9501.htm>
- DARPA Neural Network Study*. (1988), AFCEA International Press, p. 60
- Fayyad, U., Piatetsky-Shapira, G., Smyth, P., and Uthurusamy, R., eds. (1996). *Advances in Knowledge Discovery and Data Mining*. Cambridge, MA: MIT Press.
- Flaherty, C., and Pateerson, D. (2002). Predicting Child Physical Abuse Recurrence : Comparison of Neural Network to Logistic Regression. Retrieved 9 April 2005 from <http://www2.uta.edu/cussu/husita/prposals/flaherty.htm>
- Fletcher, D., and Goss, E. (1993). Forecasting with Neural Networks: An Application using Bankruptcy Data, *Information and Management*, Vol. 24, No. 3, pp. 159- 167.
- Gorr, W.L., Nagin, D., and Szczypula, J. (1994). Comparative Study of Artificial Neural Network and Statistical Models for Predicting Student Grade Point Averages, *International Journal of Forecasting*, Vol. 10, No. 1, pp. 17-34.
- Grossman. R., Kasif, S., Moore, R., Rocke, D., and Ullman, J. (1998). Data Mining Research : Opportunities and Challenges. Retrieved 9 April 2005 from <http://www.rgrossman.com/epapers/dmr-v8-4-5.htm>
- Han, J., and Kamber, M. (2001). *Data Minig : Concepts and Techniques*. Morgan Kaufmann, New York
- Hardgrave, B.C., Wilson, R.L., and Walstrom, K.A. (1994). Predicting Graduate Student Success: A Comparison of Neural Networks and Traditional Techniques, *Computers and Operations Research* , Vol. 21, No. 3, pp. 249-263
- Hayashi, Y., Setiono, R., and Yoshida, K. (2000). A Comparison between Two Neural Network Rule Extraction Techniques for the Diagnosis of Hepatobiliary Disorders. Retrieved 12 April 2005 from <http://citeseer.ist.psu.edu/cache/papers/cs/15007/http:zSzzSzwww.comp.nus.edu.sgzSz~rdyszSzcomp-hepar.pdf/hayashi00comparison.pdf>
- Haykin, S. (1994), *Neural Networks: A Comprehensive Foundation*, NY: Macmillan, p. 2

- Hong, S. J., and Weiss, S. M. (1999). *Advances in Predictive Model Generation for Data Mining*. Retrieved 9 April 2005 from [http://citeseer.ist.psu.edu/cache/papers/cs/10282/http:zSzzSzwww.research.ibm.comzSzarzSzpaperszSzpdfzShongweiss99\\_with\\_cover.pdf/hong99advances.pdf](http://citeseer.ist.psu.edu/cache/papers/cs/10282/http:zSzzSzwww.research.ibm.comzSzarzSzpaperszSzpdfzShongweiss99_with_cover.pdf/hong99advances.pdf)
- Kontkanen, P., Myllymaki, P., and Tirri, H. (1996). *Predictive Data Mining with Finite Mixtures*. Retrieved 9 April 2005 from <http://citeseer.ist.psu.edu/cache/papers/cs/2771/http:zSzzSzwww.cs.helsinki.fiSz~tirrizSzkdd96.pdf/kontkanen96predictive.pdf>
- Marshal, S.W. (2001). *Alternative Logistic Regression : A New Method for Correlated and Longitudinal Injury*. Retrieved 9 April 2005 from [http://apha.confex.com/apha/129am/twchprogram/paper\\_26798.htm](http://apha.confex.com/apha/129am/twchprogram/paper_26798.htm)
- Pregibon, D. (1997). *Data Mining*. Statistical Computing and Graphics, 7,8. StatSoft, Inc. (2003). *Data Mining Techniques*. Retrieved 9 April 2005 from <http://www.statsoft.com/textbook/stdatmin.html>
- Ripley, B.D. (1993). *Statistical Aspects of Neural Networks, Proceedings of Invited Lectures for SemStat, Sandbjerg, Denmark*.
- Salchenberger, L. M., Cinar, E. M., and Lash, N. A. (1992). *Neural Networks: A New Tool for Predicting Thrift Failures*," *Decision Sciences*, Vol. 23, No. 4, (July/Aug.), pp. 899-916.
- Tam, K. Y., and Kiang, M. Y. (1992). *Managerial Applications of Neural Networks: The Case of Bank Failure Predictions*, *Management Science*, Vol. 38, No. 7, (July), pp. 926-947.
- Tkach, D. S. (1998). *Information Mining with the IBM Intelligent Miner Family*. Retrieved 9 April 2005 from <http://www.bios.com.au/whitefam3.pdf>
- Weiss, S., and Indurkha, N. (1998). *Predictive Data Mining: A Practical Guide*. Morgan Kaufmann, New York
- Wilson, R.L., and Sharda, R. (1994). *Bankruptcy Prediction using Neural Networks*, *Decision Support Systems*, Vol. 11, No. 5, pp. 545 - 557.
- Zurada, J.M. (1992). *Introduction To Artificial Neural Systems*, Boston: PWS Publishing Company, p. xv