

**A STUDY ON DATA MINING ALGORITHMS
FOR CHURN PREDICTION IN AN ORGANIZATION**

By

KEVIN KURNIAWAN
11302010

BACHELOR'S DEGREE
in

INFORMATION TECHNOLOGY
FACULTY OF ENGINEERING AND INFORMATION TECHNOLOGY

SWISS GERMAN UNIVERSITY

SWISS GERMAN UNIVERSITY
The Prominence Tower
Jalan Jalur Sutera Barat No. 15, Alam Sutera
Tangerang, Banten 15143 - Indonesia

July 2017

Revision after the Thesis Defense on 20th July 2017

STATEMENT BY THE AUTHOR

I hereby declare that this submission is my own work and to the best of my knowledge, it contains no material previously published or written by another person, nor material which to a substantial extent has been accepted for the award of any other degree or diploma at any educational institution, except where due acknowledgement is made in the thesis.

Kevin Kurniawan

Student

Date

Approved by:

Alva Erwin, ST, M.Sc., MTI

Thesis Advisor

Date

Dipl. -Inf. Kho I Eng

Thesis Co-Advisor

Date

Dr. Ir. Gembong Baskoro, M.Sc

Dean

Date

Kevin Kurniawan

ABSTRACT

A STUDY ON DATA MINING ALGORITHMS FOR CHURN PREDICTION IN AN ORGANIZATION

By

Kevin Kurniawan
Alva Erwin, ST, M.Sc., MTI, Advisor
Dipl. -Inf. Kho I Eng, Co-Advisor

SWISS GERMAN UNIVERSITY

Employee churn and customer churn are two common problems for an organization as it indicates organization loss of resources and profits. Hence, implementing churn prediction helps to minimize those problems by detect potential churn employees or customers and apply targeted strategies to prevent them churning. By minimizing churn rate could maximize organizations revenue and productivity. This research presents churn prediction for human resource as a tool to prevent loss of valuable employees and for marketing as a tool to prevent customers' loss in telecom industry. The author compared three common data mining algorithms on both datasets in predicting churn. The experiment result shows the overview prediction performance of each data mining algorithms in different unit of analysis. Overall, the best prediction was performed by Random Forest with PCC score 85.9% in employee churn and 95.4% in customer churn, F-Measure score 26.2% in employee churn and 82.2% in customer churn. However, Random Forest's F-Measure score in employee was not the highest as Neural Network has F-Measure score 47.1%. From this research, it can be concluded that Random Forest is a good data mining algorithm to perform churn prediction in different industries. Moreover, more comparison is recommended for real implementation in order to achieve best prediction model.

Keywords: customer churn, employee churn, data mining, churn prediction, telecom, Random Forest



DEDICATION

I dedicate this works for my beloved family, friends, and country: Indonesia.



ACKNOWLEDGEMENTS

As completion of this thesis work, I would like to express my gratitude for several parties involved:

Swanly Kurniawan & Lilywaty Utama - as parents, they motivated me throughout the process of making. In addition, they supported me financially to make this thesis work done.

Alva Erwin and Kho I Eng - as the advisor and co-advisor of my thesis work. They helped to shape the ideas of the thesis. From the starting point, they had confidence on me to accomplish the thesis topics. Besides, they supported me to analyzed problems encountered during the process, problem solving, and directed me to think critically.

Swiss German University - as the organization who provided me this opportunity to create thesis work and arranged all the needs in administration and thesis defense.

Finally, I also appreciated the support from my colleagues as they had being supportive and giving critics and ideas throughout the process.

SWISS GERMAN UNIVERSITY

TABLE OF CONTENTS

	Page
STATEMENT BY THE AUTHOR.....	2
ABSTRACT	3
DEDICATION	5
ACKNOWLEDGEMENTS	6
TABLE OF CONTENTS	7
LIST OF FIGURES.....	10
LIST OF TABLES.....	11
CHAPTER 1 - INTRODUCTION	12
1.1 Background.....	12
1.2 Research Problems	13
1.3 Research Objectives	14
1.4 Significance of Study.....	14
1.5 Research Questions.....	15
1.6 Hypothesis	15
CHAPTER 2 - LITERATURE REVIEW	16
2.1 Employee Turnover as Churn: Definition and concepts	16
2.2 Customer Churn Management in Telecom.....	16
2.3 Data Mining.....	17
2.4 Data Mining Techniques	17
2.5 Data-mining Algorithms for Churn Prediction.....	18
2.5.1 Random Forest.....	18
2.5.2 Support Vector Machine.....	20
2.5.3 Artificial Neural Network.....	20
2.6 Variable (Feature) Importance.....	22
2.7 Feature Selection	22
2.7.1 Embedded approaches	23
2.7.2 Filter approaches.....	23
2.7.3 Wrapper approaches	23

2.8	Pearson's Correlation Coefficient Matrix.....	23
2.9	Data Source.....	23
2.9.1	Employee Turnover Dataset.....	24
2.9.2	Telecom Churn Dataset.....	26
2.9.3	Relationship of Selected Datasets.....	28
2.10	Related Work.....	29
2.10.1	Contribution.....	33
CHAPTER 3 - RESEARCH METHODS.....		34
3.1	Scope of Research.....	34
3.2	Research Limitation.....	34
3.3	Research Process.....	34
3.3.1	Data Collection.....	35
3.3.2	Data Preprocessing Phase.....	36
3.3.3	Data Validation.....	36
3.3.4	Churn Prediction.....	37
3.3.5	Performance Measurement.....	37
3.3.6	Comparison & Result Analysis.....	39
CHAPTER 4 - RESULTS AND DISCUSSIONS.....		40
4.1	Experimental Setup.....	40
4.1.1	Hardware Setup.....	40
4.1.2	System Architecture.....	40
4.2	Parameter Tuning of Prediction Model.....	41
4.2.1	Support Vector Machine.....	41
4.2.2	Random Forest.....	42
4.2.3	Neural Network.....	43
4.3	Result Performance.....	44
4.3.1	Employee Dataset.....	44
4.3.2	Telecom Churn Dataset.....	46
4.4	Feature Importance.....	47
4.4.1	Employee Dataset.....	48
4.4.2	Telecom Dataset.....	48
4.5	Discussion.....	49

4.5.1 Employee Dataset	50
4.5.2 Telecom Dataset	51
CHAPTER 5 - CONCLUSION AND RECOMMENDATION	53
5.1 Conclusions	53
5.2 Recommendations	54
APPENDIX A.....	55
APPENDIX B.....	57
APPENDIX C.....	63
REFERENCES	68
CURRICULUM VITAE	73

