Bank Customer Churn Prediction Based On Transaction Behavior

Fernando GPR

179460U

Dissertation submitted to the Faculty of Information Technology,University of Moratuwa, Sri Lanka for the partial fulfillment of therequirements of the Honours Degree of Bachelor of Science inInformation Technology.

2020

Declaration

We declare that is our own work and has not been submitted in any form for another degree or diploma at any university or other institution of tertiary education. Information derived from the published or unpublished work of others has beenacknowledged in the text and a list of references is given.

Name of Student — Signature of Student

G.P.R Fernando ……

Date: ……..

Supervised by

Name of Supervisor — Signature of Supervisor

Mr. ChamanWijesiriwardana ……

Date:……...

Acknowledgements

I would first like to thank my supervisor, Mr.ChamanWijesiriwardanaof the Senior Lecturer, Faculty of Information Technology, University of Moratuwa. He consistently allowed this thesis to be my own work, but steered me in the right the direction whenever he thought I needed it.his guidance, supervision, advices and sparing valuable time thorough the research project.

I would also like to thank the experts who were involved in the data extraction for this research project. Without their passionate participation and input, the extraction could not have been successfully conducted.

Finally, I must express my very profound gratitude to my parents and to my spouse for providing me with unfailing support and continuous encouragement throughout my years of study and through the process of researching and writing this thesis. This accomplishment would not have been possible without them.

Abstract

Customer churn has become a huge problem in many banks because it costs a lot to acquire a new customer than retaining an existing one. Possible churners in a bank can be identified with the use of a customer churn prediction model and as a result the bank can take necessary actions to prevent those customers from leaving the bank. In order to set up such a model in a bank, few things have to be considered such as how a churner in a bank is defined and which variables and methods should be used. This proposes that a churner for that bank should be defined as a customer who has not been active for the last three months as per the bank's definition of an active customer. Behavioral and demographic variables should be used as an input for the model and classification should be used as a technique.

Table of Contents

Introduction 1
1.1 Background 2
1.2 Problem 3
1.3 Aim and Objectives 4
1.4 The scope of research 4
1.5 Summary 4
Literature Review of Bank Customer Churn Prediction 5
2.1 Introduction 5
2.2 Data Mining Techniques 5
2.2.1 Decision Tree 6
2.2.2 Neural Network 7
2.2.3 Naïve Bayes 7
2.2.4 Support Vector Machine 8
2.3 Literature Review 8
2.4 Summary 10
Technologies Adapted 11
3.1 Introduction 11
3.2 Data Mining Tools 11
3.2.1 Rapid Miner 11
3.2.2 Orange 11
3.2.3 WEKA 12
3.3 Java 12
3.4 MySQL 12
3.4 Summary 13
A Novel Approach to Predict Banking Churn Customer 14
4.1 Introduction 14
4.2 Hypothesis 14
4.3 Methodology 15
4.3.1 Data Collection 15
4.3.2 Data Preprocessing 15

4.3.3 Feature Extraction 15
4.3.4 Classification 15
4.4 Input 16
4.5 Output 18
4.6 Process 18
4.7 Features 18
4.9 Summary 19
5.1 Introduction 20
5.2 Top level architecture 20
5.3 Data set 21
5.4 Database 21
5.4.1 Staging Data 21
5.4.2 Data Preparation 22
5.5 Churn-LIB 22
5.6 Data Preprocessing 23
5.7 Exploring Dataset 24
5.8 Classification Model Building 24
5.9 Summary 24
Implementation 25
6.1 Introduction 25
6.2 Data Transformation 25
6.3 Explorer the Data 27
6.4 Preprocessing 30
6.5 Feature selection 31
6.6 Selecting the Algorithms for customer churn prediction 36
6.7 Summary 37
Results and evaluation 38
7.1 Introduction 38
7.2 Evaluation of classification techniques 38
7.3 Ensemble learning 40
7.5 Discussion on model selection 43
7.5.1 Confusion Matrix evaluation 43

7.5.2 Sensitivity and Specificity evaluation 45
7.6 Summary 47
Conclusion and Further Work 48
8.1 Introduction 48
8.2 Bank Churn Predictor Conclusion 48
9.2 Limitations 49
9.2 Future Developments 49
9.4 Summary 49
References 50
Appendix - A 52
Appendix - B 58
Appendix - C 61

LIST OF FIGURES

Figure 1: Distribution of data mining techniques 6
Figure 2 : Model building methodology 16
Figure 3 : Customer master data 16
Figure 4 : Transaction data 17
Figure 5 : Churn prediction design 20
Figure 6 : Records staging tables 22
Figure 7 : Churn-LIB application 23
Figure 8 : Customer master – transaction data files 25
Figure 9 : Transaction Data Transform 26
Figure 10 : Statistical logic in churn-LIB 27
Figure 11: Feature list 28
Figure 12 : Replace missing values 31
Figure 13 : PCA value decomposition 32
Figure 14 : WEKA Principal components evaluator 33
Figure 15: PCA results combination of features 33
Figure 16 : PCA Resulting features 34
Figure 17 : Precision and recall 38
Figure 18 : WEKA experimenter test output 43
Appendix A19:1 BayesNet Model Accuracy 52
Appendix A20:2 BayesNet Model Evaluation on test set 52
Appendix A21:3 Naïve Bayes Model Accuracy 53
Appendix A22: 4 Naïve Bayes Model Evaluation on test set 53
Appendix A23: 5 J48 Model Accuracy 54
Appendix A24: 6 J48Model Test Evaluation on test set 54
Appendix A25: 7 Support Vector SMO Model Accuracy 55
Appendix A26: 8 Support Vector SMO Evaluation on test set 55
Appendix A27:9 RandomForest Model Accuracy 56
Appendix A28:10 RandomForest Model Evaluation on test set 56
Appendix A29:11 Bagging Model Accuracy 57
Appendix A30:12 Bagging Model Evaluation on test set 57
Appendix B31:1 Bagging J48 58
Appendix B32:2 Bagging SMO 58
Appendix B33:3 Randomization 59
Appendix B34:4 AdaBoostM1 for classifier J48 59
Appendix B35: 5 AdaBoostM1 for classifier SMO 60
Appendix B36:6 WEKA experimenter 61

LIST OF TABLES

Table 1 : PCA Feature Selection 35
Table 2 : Classification Evaluation Measurements 39
Table 3 : Classification technique result comparison 39
Table 4 : Evaluation on test set result comparison 40
Table 5 : Ensemble learning result comparison 41