

**CUSTOMER CHURN REASONING ANALYSIS MODEL
FOR TELECOMMUNICATION INDUSTRY.**

Kapugama Geeganage Kasun Thilina

198777H

Faculty of Information Technology

University of Moratuwa

July 2022

**CUSTOMER CHURN REASONING ANALYSIS MODEL
FOR TELECOMMUNICATION INDUSTRY.**

Kapugama Geeganage Kasun Thilina

198777H

Dissertation submitted to the Faculty of Information Technology,

University of Moratuwa, Sri Lanka

for the partial fulfillment of the requirements of the

MSC in Information Technology

July 2022

Declaration

I have used my own work for preparing this report. All the other published and non-published materials which were used in the research study have been acknowledged in the references section on the thesis.

K G Kasun Thilina

Signature of Student:

Date:

Supervised By

Dr. C.P. Wijesiriwardena

Signature of Supervisor:

Lecturer

Faculty of Information Technology

University of Moratuwa

Date:

Acknowledgements

I am so grateful to Dr. Chaman Wijesiriwardana, Senior Lecturer, Faculty of Information Technology, and University of Moratuwa. My research supervisor for guiding me in all aspects and supported in achieving the objectives and milestones in order to complete the project successfully.

My next most gratitude pays to all academic staff from the Faculty of Information and technology, who shared their vast knowledge throughout these two years by providing me with a good environment, which influenced a lot to achieve this goal.

Then, I would thank all non-academic staff of the University of Moratuwa, Sri Lanka for all the efforts and facilities that it has taken to contribute towards this postgraduate program and all my colleagues for their support to complete this research study in many ways. All those whom I have not mentioned name-wise who helped and motivated me.

Furthermore, I extend my earnest thanks to my parents, and siblings, for always being my support system and motivating me always.

Abstract

Customer churn is the most impactful problem in every business and industry. Therefore, every company tries their best to satisfy and maintain existing customers. Today telecommunications companies are facing this problem frequently due to increasing demand of customers every day. It is very difficult to gather new customers and need to allocate a huge cost from company revenues to acquire new customers compared to retaining the existing customer, therefore it is more important to increase their customer retention and work for that. This research is based on churn customer information and the primary objective of this research is to predict the churn reason of a given customer who has predicted to be churn using modern data analytics techniques. It include Logistic Regression, Naive Bayes, Random Forest, Decision Tree, K-Nearest Neighbor, Support Vector Machine and Gradient Boost Classifier. Further, Hybrid Model has been considered using Voting Classifier ML model. The dataset used in this research is obtained through the Data Warehouse of one of the leading telecommunication companies.

Table Contents

Declaration.....	i
Acknowledgements.....	ii
Abstract.....	iii
Table Contents.....	iv
Abbreviations.....	vii
List of Figures.....	viii
List of Tables.....	ix
Chapter 01 Introduction.....	1
1.1 Background and Motivation.....	1
1.2 Aim and Objectives.....	2
1.2.1 Aim.....	2
1.2.2 Objectives.....	2
1.3 Problem Definition.....	2
1.4 Problem Statement.....	2
1.5 The Way of Proceeding.....	3
1.6 Report Outline.....	3
1.7 Summary.....	3
Chapter 02 Customer Churn Analysis Literature Review.....	4
2.1 Introduction.....	4
2.2 Rule Based approach of Customer Churn Prediction.....	4
2.3 Data Mining based customer segmentation.....	6
2.4 Customer Churn Prediction Models.....	9
2.5 Summary.....	11
Chapter 03 Technology Adapted.....	12
3.1 Introduction.....	12
3.2 Involvement of Data Mining.....	12
3.2.1 Data Mining Algorithms and Techniques.....	12
3.3 Model Development Programming Languages and Tools.....	14
3.3.1 Programming Languages.....	15
3.3.2 Tools.....	15
3.4 Data Collection Techniques.....	15
3.5 Summary.....	16
Chapter 04 Churn Reasoning Analysis Approach.....	17
4.1 Introduction.....	17

4.2 Input.....	17
4.3 Output	17
4.4 Process	17
4.4.1 Data Selection	18
4.4.2 Data Preprocessing.....	18
4.4.3 Data Mining	18
4.4.4 Evaluation	18
4.5 Features.....	19
4.6 Summary.....	20
Chapter 05 Analysis and Design	21
5.1 Introduction	21
5.2 Top Level Design of the Proposed Classification Model.....	21
5.2.1 Data Gathering	21
5.2.2 Data Preprocessing.....	22
5.2.3 Modeling.....	23
5.3 Summary.....	24
Chapter 06 Implementation of the Study.....	25
6.1 Introduction	25
6.2 Data Gathering.....	25
6.2.1 Loading Data.....	25
6.3 Data Analysis and Visualization.....	25
6.4 Data Preprocessing	30
6.5 Checking for Missing Values	30
6.5.1 Remove Outliers	30
6.5.2 Categorical to Numerical Conversion.....	31
6.5.3 Class Imbalance to Balance	32
6.5.4 Applying Appropriate Classification Technique	34
6.6 Summary.....	34
Chapter 07 Evaluation	35
7.1 Introduction	35
7.2 Evaluation for Classification	35
7.2.1 Mean Score	35
7.2.2 Confusion Matrix	36
7.2.3 Feature Importance	37
7.3 Summary.....	39

Chapter 08 Conclusion and Future Work	40
8.1 Introduction	40
8.2 Conclusion	40
8.3 Limitation	41
8.4 Future Developments.....	41
8.5 Summary.....	41
Reference	42
APPENDIX A.....	44
APPENDIX B	47
APPENDIX C	51

Abbreviations

DM	Data Mining
WEKA	Waikato Environment for Knowledge Analysis
CRL	Classification by Rule Learning
DMEL	Data Mining by Evolutionary Learning
EA	Exhaustive Algorithm
GA	Genetic Algorithm
CA	Covering Algorithm
LA	LEM2 Algorithm
DT	Decision Tree
ANN	Artificial Neural Networks
KNN	K-Nearest Neighbors
SVM	Support Vector Machine
TP	True Positive
TN	True Negative
FP	False Positive
FN	False Negative
IQR	Interquartile Range
SMOTE	Synthetic Minority Oversampling Technique
CSV	Comma Separated Values

List of Figures

Figure 2.1 Proposed Methodology [9]	5
Figure 2.2 Hybrid methodology Recall values [1].....	7
Figure 2.3 Hybrid methodology Precision values [1].....	7
Figure 2.4 Proposed Methodology for Hybrid Classifier [1].....	7
Figure 2.5 Process of Predictive Model Creation [2]	8
Figure 2.6 Churn Prediction Framework	9
Figure 2.7 Analysis of outcomes in data mining techniques [11].....	10
Figure 5.1 Classification Model.....	21
Figure 5.2 Data Source	21
Figure 5.3 K-fold Cross-Validation Technique	24
Figure 6.1 The CSV Data File Import	25
Figure 6.2 The Statistics of Numeric Variables.....	25
Figure 6.3 The Churn Reason Frequency	26
Figure 6.4 Analysis of Gender	27
Figure 6.5 Analysis of Senior Citizen.....	27
Figure 6.6 Analysis of Current Package	28
Figure 6.7 Analysis of Payment Method	28
Figure 6.8 Analysis of Distribution Curves of Average Voice Monthly Charges.....	29
Figure 6.9 Analysis of Distribution Curves of Average Data Monthly Charges.....	29
Figure 6.10 Result of Missing Values.....	30
Figure 6.11 IQR Technique Summary.....	31
Figure 6.12 Categorical Variables	31
Figure 6.13 Numerical Conversion Summary of Dataset.....	32
Figure 6.14 X and y Variables Initialized.....	32
Figure 6.15 Class Imbalance Behavior	33
Figure 6.16 Balanced Class Variable Summary	33
Figure 7.1 Classification Performance Graphical Mean Scores	36
Figure 7.2 Feature Importance Graphical Scores for Random Forest	38
Figure 7.3 Feature Importance Graphical Scores for Decision Tree	38

List of Tables

Table 2.1 Recall and F-Measure Results	11
Table 4.1 Feature	20
Table 5.1 Churn Customer Variables and Data Types	23
Table 6.1 Numeric Variables Description	26
Table 7.1 Classification Performance Numerical Mean Scores	35
Table 7.2 Evaluation Measurements for Classifiers	36
Table 7.3 Classification Performance Numerical Scores Based on Confusion Matrix	37
Table 7.4 Feature Importance Numerical Scores.....	38