Level 4

DATA MINING TECHNIQUES TO IDENTIFY FRAUDS IN WATER BOTTLE DELIVERY AND PREDICT THE FUTURE DEMAND FOR SALES TRENDS

D.A.S.D Kalansuriya

169314T

Supervised by:

Mr. Saminda Premaratne

(Senior Lecturer)

Department of Information Technology

University of Moratuwa

2018

Declaration

I declare that this thesis is my 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 Information derived from the published or unpublished work of others has been acknowledged in the text and a list of references is given.

Name of Student
D.A.S.D Kalansuriya

Signature of Student
Date: ……

Supervised by

Name of Supervisor
S. C. Premaratne

Signature of Supervisor
Date: ……

Acknowledgment

I would like to express my deepest Gratitude to my project supervisor Mr. S.C. Premaratne for his patient guidance, enthusiastic encouragement and useful reviews to success this project. Furthermore, my next big thank goes to Dr. Mohamed Firdhous who taught us Research Methodology and Dr Mohamed Firdhous.

Moreover, I would also like to acknowledge with much appreciation the help of the all the lecturers in M.Sc. in Information Technology degree program of Faculty of IT, who gave their fullest support success this program, by sharpen our knowledge and ideas throughout these two years as they were the illumination which lit up our pathways to success.

My special thanks should go to Director at American Premium Water System (Pvt) Ltd Mr. Fayaz Fazal for giving me details, which are helpful to complete the report.

Apart from the people who were directly involved, many more helped to make this project a success.

Finally, I wish to thank to my family and friends for their support and encouragement throughout my study

Abstract

Data mining is a subset of databases management and it mainly applicable to large and complex databases to eliminate the randomness and discover the hidden pattern. Fraud detection in data mining is the process of identifying fraudulent acts by analyzing the dataset. Research is based on identifying fraudulent acts of water bottle delivery process. The research study focusses on to manage the invoicing process with the water delivery process. Due inefficacies in the water delivering process bottle lost cost in the last six months is Rs 213,070.00 approx. Through detecting fraudulent acts, the institutes can save resources and cost [3], for this study a sample data set has been used to identify how the fraudulent activities are occurring. Sample dataset has been selected from where data entry person had found physical evidence that the bottle had been sold for outsiders.

Data mining tools which used to detect frauds are Naïve Bayes, Decision Trees, and neural networks. By developing predictive models can be generated to estimate things such as the probability of fraudulent behavior. ROC curves have deployed for model assessment to provide a more intuitive analysis of the models and confusion matrix is has used to describe the performance of a classification model on the test data for which the true values are known.

Contents

Declaration ii
Acknowledgment iii
Abstract iv
List of Figures viii
List of Tables ix
List of Equation x
List of Abbreviation xi
1 Introduction 1
1.1 Aim 3
1.2 Objectives 3
1.3 Assumption 3
1.4 Thesis structure 3
2 Literature Review 5
2.1 Introduction 5
2.1.1 Data Mining Methodology 5
2.1.2 Standard Data Mining Process 6
2.1.3 Data Mining Methods 7
2.1.4 Data Mining Techniques 8
2.2 Background to Frauds 9
2.2.1 Related Works 10
2.2.2 Fraud Detecting Methods 12
2.3 Summary 12
3 Technology Adopted 13
3.1 Introduction 13
3.2 Selected methods or techniques 13
3.3 Tools using for a data mining 16
3.4 Summary 17
4 Analysis and the Design 18
4.1 Introduction 18
4.2 Attributes of the analysis 18
4.3 Sample Selection Process 19
4.4 Summary 20
5 Implementation 21
5.1 Introduction 21

5.2 Data collection....21
5.3 Data Preparation....21
5.3.1 Customer Selection....21
5.3.2 Consumption levels....22
5.3.3 Customer Complaints....23
5.3.4 Stock available....24
5.3.5 Missed delivery....25
5.3.6 Housed closed....25
5.3.7 Instances of manual tickets....26
5.3.8 Manual Invoices....27
5.4 Consumption Predication Methods....28
5.4.1 Naïve Bayes....28
5.4.2 Decision Tree....30
5.4.3 ANN (Neural Networks)....32
5.4.4 Accuracy of the model....41
5.5 Possible Fraud detection....41
5.5.1 Naïve Bayes....42
5.5.2 Decision Tree....45
5.5.3 Neural Networks....47
5.5.4 Accuracy of the Algorithms....49
5.6 Summary....50
5.6.1 Consumption Prediction....50
5.6.2 Fraud Detection....50
6 Implementation of the Model....51
6.1 Introduction....51
6.2 Result evaluation....51
6.2.1 Consumption Prediction Model Evaluation....51
6.2.2 Fraud Detection Model....53
6.3 Summary....55
7 Discussion....56
7.1 Introduction....56
7.2 Importance of the research....56
7.3 Future Works....59
7.3.1 Areas of future study....59
8 Reference....60

9 Appendix 63
9.1 Code snippet to generate the summary from R studio 63
9.2 Code Snippet for data visualize as Bar Chart 63
9.3 Model Development Process 64
9.3.1 Model Selection 64
9.3.2 Saving Model 64
9.3.3 Model Evaluation 64
9.3.4 Attribute selection 65
9.3.5 Confusion Matrix 65

List of Figures

Figure 2.1.1-1: Revenue Loss Forecast....1
Figure 2.2.2-1:Research Methodology 19
Figure 2.2.2-1:Location wise summary 20
Figure 5.4.1-1:Naïve Bayes Summary Window 28
Figure 5.4.1-2: Naive Bayes Model....29
Figure 5.4.2-1:Decsion Tree Summary indow....30
Figure 5.4.2-2: Desicion Tree model 31
Figure 5.4.2-3:Tree View....32
Figure 5.4.3-1:Neural Network model....35
Figure 5.4.3-2:Four, Four Two Layer....41
Figure 5.5.1-1:Naive Bayes model 44
Figure 5.5.2-1:Decision tree result window....45
Figure 5.5.2-2:Desion Modeler....46
Figure 5.5.2-3:Decision tree view....47
Figure 5.5.3-1:Two nodes, one layer 49
Figure 6.2.1-1:Results of the model selection 51
Figure 6.2.1-2:Knowledge flow Steps 51
Figure 6.2.1-3:Attribute selection window 52
Figure 6.2.2-1:Fraud detection Model 53
Figure 6.2.2-2:Attribute selection window 54
Figure 9.3.5-1:TRP and FPR[33]....65

List of Tables

Table 2.2.2-1:Ranking table of the Consumption Problem 19
Table 5.3.1-1:Customer Categorization.....22
Table 5.3.2-1:Class Lables of Water Consumption.....22
Table 5.3.3-1:Count of Complaints 23
Table 5.3.3-2:Complaint data selection 24
Table 5.3.4-1:Stock data selection.....24
Table 5.3.5-1:Missed Delivery data selection.....25
Table 5.3.6-1:House closed data selection.....26
Table 5.3.7-1:Manual tickets data selection 26
Table 5.3.8-1:Manual Invoice data selection.....27
Table5.4.3-1: Neral Natwork Result.....40
Table 5.4.4-1:Acuracy table.....41
Table 5.4.4-1:Rule Based to classify 42
Table 5.5.4-1:Acuracy table.....49
Table 6.2.1-1:Result table of Consumption predicion.....52
Table 6.2.1-2:Attribute raninking table 52
Table 6.2.2-1:Result table of Fraud detection.....54
Table 6.2.2-1:Classifications table.....56
Table 6.2.2-2:Detailed Accuracy by Class (Nureal Networks) 57

List of Equation

Equation 3.2-1:Entropy.....13
Equation 3.2-2:"Entropy" for the target given a bin.....14
Equation 3.2-3:Information Gain.....14
Equation 3.2-4:Naive Bayes 14
Equation 3.2-5: Decision Tree Algorithm 15
Equation 3.2-6:Information needed 15
Equation 3.2-7:Information gained.....15
Equation 3.2-9:Backpropagation: A neural network learning algorithm[26].....16

List of Abbreviation

ANN	Neural Networks
ROC	Receiver operating characteristic