

ANALYZING CUSTOMER BEHAVIOURS AND PREDICTING AN OPTIMAL CREDIT LIMIT

H. M. Methma Tharindi Bandara

(219317L)

Degree of Master of Science in Computer Science

Department of Computer Science and Engineering

University of Moratuwa

Sri Lanka

April 2023

ANALYZING CUSTOMER BEHAVIOURS AND PREDICTING AN OPTIMAL CREDIT LIMIT

H. M. Methma Tharindi Bandara
(219317L)

Thesis/Dissertation submitted in partial fulfillment of the requirements for the
degree

Master of Science in Computer Science specialization in Data Science

Department of Computer Science and Engineering

University of Moratuwa

Sri Lanka

April 2023

DECLARATION

I declare that this is my own work, and this thesis/dissertation does not incorporate without acknowledgement any material previously submitted for a degree or diploma in any other University or Institute of higher learning and to the best of my knowledge and belief it does not contain any material previously published or written by another person except where the acknowledgement is made in the text. I retain the right to use this content in whole or part in future works (such as articles or books).

Signature:

Date: 2023-08-07

The supervisor/s should certify the thesis/dissertation with the following declaration.

The above candidate has carried out research for the PhD/MPhil/Masters thesis/dissertation under my supervision. I confirm that the declaration made above by the student is true and correct.

Name of the supervisor: Dr. Buddhika Karunaratne

Signature of the supervisor:

Date: 2023-08-07

ACKNOWLEDGEMENT

I want to express my deep and sincere gratitude to Dr. Buddhika Karunaratne for guiding me throughout the research with continued support and encouragement. His supervision immensely helped me in setting goals and engaging in the study in a timely manner.

I want to express my most incredible gratitude to the Department of Computer Science and Engineering, the University of Moratuwa, for providing the support to overcome this effort. Finally, my heartfelt gratitude goes to my parents who supported me throughout this endeavour.

ABSTRACT

Analyzing customer behaviours to gain many vital insights is a prominent topic in different industries. The main reason for such analysis is that it helps the businesses to identify valuable insights that could uplift both company profit and customer satisfaction. This research focuses on such area where both customer satisfaction and company profit could be uplift by identifying customer behaviours. The selected customer base is an LTE broadband customer base of a telecommunication company. The objective of this research is to predict an optimal credit limit for customers by analyzing different customer behaviours benefiting both the customer and the company. In order to identify different customer behaviors in terms of their payment and data usage patterns, clustering algorithms were used. This research discusses different cluster quality indexes to measure the goodness of the clusters. Once the customers are clustered into different groups, regression models were used to derive customized formular to identify the relationship between the profit generation, credit limit and the other features. Then dynamic programming is used to identify the optimal credit limit for each group of customers identified. Further to this a classification model is used to clarify the customers to relevant clusters identified in the future. The outcome of this research shows two different customer behaviours resulted with two different credit limits.

Keywords: Customer segmentation, Clustering, Regression, Classification, dynamic programming

Table of Contents

DECLARATION	3
ACKNOWLEDGEMENT	4
ABSTRACT	5
LIST OF FIGURES	8
LIST OF TABLES	8
LIST OF ABBREVIATIONS	9
1.0 INTRODUCTION	10
1.1 Research Problem	10
1.2 Research Objective	11
2.0 LITERATURE REVIEW	12
4.1 Customer segmentation	12
4.2 Optimization Techniques	16
3.0 REASEARCH GAP	20
4.0 METHODOLOGY	21
4.1 System Architecture	21
4.2 Data set	21
4.3 Data Preprocessing	22
4.4 Dimension reduction	22
4.5 Outlier Detection	22
4.6 Clustering algorithms	22
4.6.1 SOM – Self Organizing Maps	23
4.6.2 K-Means	24
4.6.3 Hierarchical clustering	24
4.6.4 Density Based Clustering	25
4.7 Supervised Learning	25

4.7.1 Regression model and optimization	25
4.7.2 Classification Models	27
4.8 Implementation	29
5.0 Evaluation	30
5.1. Cluster Quality evaluation	30
5.1.1 Silhouette score	30
5.1.2 Calinski_harabasz_score	30
5.1.3 Davies Bouldin Score	31
5.2 Classification Model Evaluation	31
7.0 RESULTS	33
7.1 Dimension reduction	33
7.2 Outlier detection	33
7.3 Clustering	35
7.3.1 Hierarchical Clustering Result	36
7.3.2 DBSCAN Clustering	37
7.3.4 K-Means	38
7.4 Regression model	38
7.5 Optimization	39
7.6 Classification	39
7.7 Discussion	41
CONCLUSION	44
REFERENCES	45

LIST OF FIGURES

Fig.4.1 System Architecture

Fig.4.2 Clustering model steps

Fig.5.1 Confusion matrix

Fig.7.1 PCA components vs cumulative variance graph

Fig.7.2 SOM distance Map for outlier detection after hyper parameter tuning

Fig 7.3 Quantization Error Distribution Boxplot

Fig.7.4 PCA components vs cumulative variance graph

Fig.7.5 SOM distance Map for node weights clustering

Fig7.6 Dendrogram

Fig7.7 K- Distance graph

Fig7.8 Feature Correlation of attributes of model 1.

Fig7.9 Feature Correlation of attributes of model 2.

Fig 7.10 Actual and the predicted profit for each class 0

Fig 7.11 Actual and the predicted profit for each class 1

Fig7.12 Predicted profit for each customer category.

Fig 7.13 Customer status of clusters

LIST OF TABLES

Table 7.1 Hyper parameter tuning for SOM.

Table 7.2 Hyper parameter tuning for SOM.

Table 7.3 Cluster quality measure values for hierarchical clustering

Table 7.4 Cluster Quality Scores for different epsilon values of DBSCAN algorithm

Table 7.5 Cluster Quality Scores for K-Means algorithm

Table 7.7 Classification Model accuracies

Table 7.8 Classification Model accuracy

Table 7.9 Classification confusion matrix

Table 7.8 Summary of Cluster results

LIST OF ABBREVIATIONS

CDR -Caller Detail Record

SOM – Self Organizing Map

SMOTE - Synthetic Minority Oversampling Technique