

LB/TH/15/2026
TH6174

**CUSTOMER SEGMENTATION AND
CROSS-SELLING RECOMMENDATION ENGINE
FOR THE TELECOM INDUSTRY: A MACHINE
LEARNING APPROACH**

Janadhi Dilmini Nanayakkara

219445C

Master of Science in Telecommunications

Department of Electronics & Telecommunications Engineering
Faculty of Engineering

University of Moratuwa
Sri Lanka

July 2025

**CUSTOMER SEGMENTATION AND
CROSS-SELLING RECOMMENDATION ENGINE
FOR THE TELECOM INDUSTRY: A MACHINE
LEARNING APPROACH**

Janadhi Dilmini Nanayakkara

219445C

Dissertation submitted in partial fulfillment of the requirements for the
degree

Master of Science in Telecommunications

Department of Electronics & Telecommunications Engineering
Faculty of Engineering

University of Moratuwa
Sri Lanka

July 2025

DECLARATION

I declare that this is my own work and this 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: 18/07/2025

The supervisor should certify the Dissertation with the following declaration.

The above candidate has carried out research for the Master of Science in Telecommunications Dissertation under my supervision. I confirm that the declaration made above by the student is true and correct.

Name of Supervisor: Prof. (Mrs) Dileeka Dias

Signature of the Supervisor:

Date:

DEDICATION

To my parents, for their endless sacrifices and unwavering belief in me, and to all those who stood by me with patience and encouragement - this achievement belongs as much to you as it does to me.

ACKNOWLEDGEMENT

This research would not have been a success without the kind support and help of many individuals and organizations. I would like to extend my sincere thanks to all of them.

First of all, I wish to express my heartfelt gratitude to my academic supervisor, Prof. (Mrs.) Dileeka Dias, for her continuous guidance, insightful feedback, and invaluable encouragement throughout the course of this research. Her guidance and dedication were crucial for this study.

I would also like to extend my special thanks to Dr. Kasun Hemachandra, the course coordinator of the PG Diploma/ MSc in Telecommunications program, for his expert advice and dedication towards mentoring students and inspiring them to achieve their full potential.

A special note of appreciation goes to all individuals who offered their assistance in gathering the required datasets and insights which contributed significantly to the successful completion of this research.

To my friends and colleagues, thank you for your encouragement, thoughtful discussions, and moral support throughout this journey.

Finally, and most importantly, I am thankful to my family for the moral support, encouragement and guidance extended in many ways at all stages of my life. This accomplishment would not have been possible without their constant support.

ABSTRACT

In the telecommunication industry, personalized service delivery and effective cross-selling strategies are vital for enhancing customer engagement and driving revenue growth. This research presents a data-driven approach to customer segmentation and personalized recommendation, aiming to promote the transition from Double-Play (Voice and Internet) to Triple-Play (Voice, Internet and IPTV) service adoption.

The study applies unsupervised machine learning techniques, namely Gaussian Mixture Models (GMM) and K-Means clustering to segment customers based on demographic, billing, and service usage attributes. Multiple internal evaluation metrics, including the Silhouette Score, Calinski-Harabasz Index, Davies-Bouldin Index and the Elbow method are employed to determine the optimal clustering structure and validate the segmentation quality.

To generate personalized service recommendations, a collaborative filtering approach based on Singular Value Decomposition (SVD) is utilized. The model predicts the most suitable IPTV package for Double-Play customers by learning from existing customer interaction patterns. The performance of the recommendation system is evaluated using Root Mean Square Error (RMSE) and Mean Absolute Error (MAE).

The implementation leverages Python-based tools and libraries for data preprocessing, modeling and visualization. The research offers a practical and scalable framework for targeted marketing and intelligent service personalization, contributing to the advancement of data-driven strategies in the telecommunications domain.

Keywords: Machine Learning, Telecommunication, Customer segmentation, Recommendation Systems

TABLE OF CONTENTS

Declaration of the Candidate & Supervisor	i
Dedication	iii
Acknowledgement	v
Abstract	vii
Table of Contents	ix
List of Figures	xiii
List of Tables	xv
List of Abbreviations	xv
List of Appendices	xix
1 Introduction	1
1.1 Background	1
1.2 Problem Statement	2
1.3 Motivation and Importance	2
1.4 Objectives and Scope	3
1.4.1 Objectives	3
1.4.2 Scope	3
1.5 Originality and Novelty	3
1.6 Thesis Outline	4
2 Literature Review	5
2.1 Data Preprocessing	5
2.2 Customer Segmentation	9
2.2.1 Overview on Clustering Techniques	9
2.2.2 Clustering Validation Indices	10
2.2.3 Existing Literature	11
2.3 Recommendation Systems	12
2.3.1 Overview on Recommendation Systems	13
2.4 Existing Research Gaps	14

2.5	Literature-Guided Selection of Clustering and Recommendation Techniques	14
2.6	Mathematical Background and Parameter Analysis of Selected Models	16
2.6.1	K-Means Clustering	16
2.6.2	Gaussian Mixture Model (GMM)	17
2.6.3	Singular Value Decomposition (SVD)	19
3	Methodology	21
3.1	Research Design Architecture	21
3.2	Design Implementation	22
3.2.1	Implementation Tools	22
3.2.2	Data Collection & Preprocessing	23
3.2.3	Customer Segmentation	29
3.2.4	Cross-Selling Recommendation Engine	32
4	Results and Discussion	35
4.1	Data Collection & Preprocessing	35
4.1.1	Dataset Merging Strategy and Selection	35
4.2	Customer Segmentation	36
4.2.1	Gaussian Mixture Model	36
4.2.2	K-Means Clustering	38
4.2.3	Model Selection for Customer Segmentation	42
4.3	Cross-Selling Recommendation Engine	45
4.3.1	Collaborative Filtering Approach	46
4.4	Verification Against Standard Benchmarks	51
4.4.1	Clustering Evaluation Benchmarks	51
4.4.2	Recommendation Evaluation Benchmarks	51
5	Conclusion and Future Work	53
5.1	Conclusion	53
5.1.1	Customer Segmentation	53
5.1.2	Cross-Selling Recommendation Engine	53
5.1.3	Business Implications	54
5.2	Future Work	54

5.3 Final Remarks	55
References	57
Appendix A Dataset Structure	61

LIST OF FIGURES

Figure	Description	Page
Figure 2.1	Categorization of Data Preprocessing Techniques (adapted from [1])	6
Figure 2.2	Data Preprocessing Taxonomy (adapted from [2])	7
Figure 2.3	Architecture of Data Preprocessing “PrePy” Package (adapted from [3])	8
Figure 2.4	Different Clustering Techniques	9
Figure 3.1	Architecture	21
Figure 4.1	Null percentage comparison across datasets	35
Figure 4.2	BIC and AIC score variation with different cluster numbers for GMM	38
Figure 4.3	Silhouette Score variation with the number of clusters for GMM	39
Figure 4.4	GMM Cluster distribution for N=8	39
Figure 4.5	GMM Cluster distribution for N=2	40
Figure 4.6	Elbow plot showing the optimal K=9 for K-Means clustering	41
Figure 4.7	Silhouette Score variation with the number of clusters for KMeans	41
Figure 4.8	K-Means cluster distribution for K=6	42
Figure 4.9	Distribution of customers across 8 clusters	44
Figure 4.10	DP/TP distribution in target clusters	45
Figure 4.11	Output for 5-fold cross validation	47
Figure 4.12	IPTV Package distribution across Triple-Play customers	48
Figure 4.13	Comparison of Actual TP Popularity vs Top-1 Recommendations (Top 15 Packages)	49
Figure 4.14	Comparison of Actual TP Popularity vs Top-1 Recommendations (Top 15 Packages) - Log Scale	49
Figure 4.15	IPTV Package distribution across Triple-Play customers	50
Figure 4.16	Customer-Package Interaction Matrix	50

LIST OF TABLES

Table	Description	Page
Table 2.1	Possible DQ checks and their preprocessing mappings	8
Table 2.2	Summary on the Literature on Customer Segmentation	12
Table 2.3	Summary on the Literature on Recommendation Systems	13
Table 3.1	Implementation Tools & Technologies	22
Table 3.2	Summary of collected datasets before preprocessing	24
Table 3.3	Summary of Datasets after preprocessing	27
Table 4.1	Some Key features on the datasets used	36
Table 4.2	Summary of Dataframes Used	37
Table 4.3	Comparison on the dataset merging strategy	38
Table 4.4	KMeans clustering performance evaluation for different K values	41
Table 4.5	Choosing the best Clustering Models	42
Table 4.6	Customer Distribution and Segment Analysis Across Clusters	45
Table 4.7	5-Fold Cross-Validation Performance	47
Table A.1	Structure of the Dataset used	61

LIST OF ABBREVIATIONS

Abbreviation	Description
AIC	Akaike Information Criterion
BIC	Bayesian Information Criterion
CH	Calinski-Harabasz
DBI	Davies-Bouldin Index
DP	Double-Play
DQ	Data Quality
GMM	Gaussian Mixture Model
LTV	Lifetime Value
ML	Machine Learning
RFM	Recency, Frequency, Monetary
RMSE	Root Mean Squared Error
SVD	Singular Value Decomposition
TP	Triple-Play
VAS	Value-Added Services

LIST OF APPENDICES

Appendix	Description	Page
Appendix -A	Dataset Structure	61