CUSTOMER LIFETIME VALUE MODELING: A MACHINE LEARNING APPROACH TO CUSTOMER SEGMENTATION BY K-MEANS AND XGBOOST

BY

Tamanna Tabassum Tithy
ID: 0242310004213003

This Report Presented in Partial Fulfillment of the Requirements for
The Degree of **Master of Science in Management & Information System**

Supervised By

Dr. Sheak Rashed Haider Noori

Professor & Head

Department of CSE

Daffodil International University

DAFFODIL INTERNATIONAL UNIVERSITY
DHAKA, BANGLADESH
JANUARY 2025

APPROVAL

This Thesis titled "**Customer Lifetime Value Modeling: A Machine Learning Approach to Customer Segmentation By K-Means and XGBoost**", submitted by Tamanna Tabassum Tithy to the Department of Computer Science & Engineering, Daffodil International University has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of MS in management & Information System and approved as to its style and contents. The presentation has been held on 11 January, 2025.

BOARD OF EXAMINERS

Dr. Sheak Rashed Haider Noori — **Chairman**
Professor & Head
Department of CSE
Faculty of Science & Information Technology
Daffodil International University

Dr. Naznin Sultana — **Internal Examiner**
Associate Professor
Department of CSE
Faculty of Science & Information Technology
Daffodil International University

Mr. Md. Sadekur Rahman — **Internal Examiner**
Assistant Professor
Department of CSE
Faculty of Science & Information Technology
Daffodil International University

Mr. Nazibur Rahman — **External Examiner**
Technical Lead - Database Administrator,
Wipro Bangladesh Telenor - Grameen Phone

DECLARATION

I hereby declare that this research has been done by me under the supervision of **Dr. Sheak Rashed Haider Noori, Professor & Head, Department of CSE,** Daffodil International University. I also declare that neither this project nor any part of this project has been submitted elsewhere for the award of any degree or diploma.

Supervised by:

Dr. Sheak Rashed Haider Noori
Professor & Head
Department of CSE
Daffodil International University

Submitted by:

Tamanna Tabassum

Tamanna Tabassum Tithy
ID: 0242310004213003
Department of MIS
Daffodil International University

ACKNOWLEDGEMENT

First and foremost, I would like to express my deepest gratitude to Almighty ALLAH for granting me the strength, wisdom, and perseverance to complete this research work.

I extend my heartfelt thanks to my supervisor, **Dr. Sheak Rashed Haider Noori, Professor & Head,** Department of CSE, Daffodil International University, Dhaka, for his invaluable guidance, support, and constructive feedback throughout the course of this research. Their encouragement and expertise have been instrumental in shaping the direction and quality of this thesis.

To my family, my parents, my siblings, especially my elder brother Fahim Shahriyer, and my Husband Sheikh Mohammad Faisal, I owe everything. Your love, prayers, and sacrifices have been my pillar of support throughout this journey.

The success of this research project would not have been possible without the collective efforts of all those involved, and I am truly thankful for their contributions.

ABSTRACT

This research utilizes advanced machine learning models to predict Customer Lifetime Value (CLV). Customer Lifetime Value (CLV) is a key business metric that estimates the total revenue a business can reasonably expect from a single customer throughout the entirety of its relationship with the company. It helps businesses understand how much each customer is worth, enabling them to make informed decisions about customer acquisition, retention strategies, and resource allocation. The research applies k-means algorithm to segment customers into distinct groups, and the XGBoost algorithm to predict CLV, offering insights into customer patterns that can enhance marketing strategies. A comparative analysis of XGBoost and K-Nearest Neighbors (K-NN) demonstrates the superior performance of XGBoost in handling complex data relationships and non-linear patterns. The results also demonstrate that using K-Means and XGBoost together makes segmentation and CLV prediction more effective, achieving a 99% classification accuracy. It provides a helpful framework for businesses to improve customer retention and profits. Moreover, focusing on ethical data usage and sustainable business practices, the research highlights the social, environmental, and ethical aspects of using machine learning in customer management.

LIST OF FIGURES

LIST OF TABLES

TABLES **PAGE NO**

TABLE OF CONTENTS

CHAPTER 1
INTRODUCTION

1.1 Introduction

In today's competitive business environment, understanding and anticipating customer behavior is essential for growth and profitability. However, focusing on every customer individually is both inefficient and costly [1]. To stay ahead of the competition, businesses are relying more and more on data to generate insights. One of the key insights is Customer Lifetime Value (CLV)- a metric which helps to predict the total revenue a company can expect from a customer over the entire duration of their relationship. By accurately predicting CLV, businesses can identify high-value customers, prioritize retention strategies, and allocate resources to maximize long-term profitability. Rather than focusing on every customer equally, businesses can concentrate on those who are most likely to bring in long-term value. This thesis explores machine learning techniques to model and forecast CLV, with a focus on customer segmentation. First, the K-Means clustering algorithm is used to segment customers into distinct groups based on their features. Once segmentation is done, the XGBoost algorithm is used to predict the lifetime value of each customer in these segments. XGBoost is particularly useful as it can handle missing data, identify non-linear relationships, and work with high-dimensional datasets with efficiency and precision [2]. By combining these two techniques, businesses can gain deeper insights needed to make more targeted marketing decisions. The main goal of this research is to create a robust, data-driven model. This enables businesses to identify high-value customers and gain a deeper understanding of their preferences. This knowledge can support personalized marketing strategies, improve customer retention, and enhance long-term profitability. Ultimately, the study aims to demonstrate how advanced machine learning techniques can improve revenue predictions and business decision-making.

1.2 Motivation

In today's market, businesses face significant challenges in understanding customer behavior and maximizing revenue. Traditional methods of customer segmentation and revenue forecasting often fail to capture the complexity of customer interactions. This limitation has created a need for more accurate and adaptable approaches.

The growing availability of customer data opens new opportunities for businesses to develop more personalized and effective strategies. However, existing methods for predicting CLV often struggle with large datasets and non-linear relationships, leading to inaccurate predictions. This is especially true in industries where customer behavior is influenced by multiple factors and is not always predictable. With data coming from various sources like e-commerce and loyalty programs, there is an increasing need for more advanced methods to accurately estimate CLV [3]. This research is motivated by the need for more accurate, adaptable, and scalable methods for predicting CLV. While machine learning techniques, such as XGBoost, have already proven effective in other predictive tasks, their potential for CLV prediction has yet to be fully realized. This study seeks to leverage XGBoost to provide businesses with a more reliable way to predict CLV, uncover key insights, and improve decision-making in areas like marketing, resource allocation, and customer retention.

1.3 Rationale of the Study

The foundation of this study focused on addressing the ongoing challenges businesses face in creating accurate, data-driven models for predicting CLV. As customer data becomes more crucial in decision-making, the need for advanced techniques that can handle large, complex datasets is growing. Traditional predictive models often fall short when it comes to understanding the complexities of customer behavior, especially with high-dimensional data where simple linear methods can't capture the interactions between various customer attributes. XGBoost offers a solid solution because it can manage these complexities, including missing data, non-linear relationships, and varying data structures. This research aims to show how XGBoost can improve CLV prediction and provide a method that can be used in different industries. By focusing on industries where customer retention is paramount, such as e-commerce, finance, and telecommunications, the research seeks to bridge the gap between theoretical advancements and practical applications [4]. This research is especially useful for industries where customer retention is essential, such as e-commerce, finance, and telecommunications. With more accurate CLV predictions, businesses can improve their marketing and retention strategies. This leads to happier customers, higher profits, and long-term success. The result will contribute to both academic research on predictive analytics and practical tools for businesses to better improve their customer relationship.

1.4 Research Questions

This thesis aims to answer the following research questions:

1. How can machine learning techniques, specifically XGBoost, improve the accuracy of CLV predictions?
2. What are the key factors that have the most impact on CLV in each dataset?
3. How can predictive CLV modeling improve customer segmentation and inform better business decisions?

1.5 Expected Output

The expected outcomes of this research are:

1. Development of an accurate CLV prediction model using XGBoost.
2. Identification of high-value customer segments for targeted marketing and retention strategies.
3. Insights into key customer features that influence CLV predictions for smarter decision-making.
4. Practical guidelines for implementing the CLV model in real-world business environments.

CHAPTER 2
BACKGROUND

2.1 Preliminaries

2.1.1 Customer Lifetime Valus (CLV): CLV is a metric used to predict how much total revenue or profit a business can expect to earn from a customer over their entire relationship. It is an important tool for managing customer relationships and helps businesses allocate their resources more effectively. According to Gupta et al., CLV can be described as the net present value of future profits from a customer [5].

2.1.2 Customer Segmentation: This is the process of dividing customers into distinct groups based on shared characteristics. It helps businesses create more personalized marketing strategies and improve how they engage with customers [6].

2.1.3 K-Means Algorithm: The K-Means algorithm is a popular clustering method that groups data into a specified number of clusters by analyzing similarities between data points. It works by minimizing the distance between each data point and the center of its assigned cluster. [7]. For example, in this study, K-Means segmented customers into four distinct groups based on features such as Lifetime Value (LTV), Customer Age, and Route Region. These clusters help businesses identify patterns and behaviors within customer groups, enabling targeted strategies

2.1.4 XGBoost (Extreme Gradient Boosting): XGBoost (Extreme Gradient Boosting) is an advanced machine learning algorithm designed to optimize the Gradient Boosting method. It combines multiple weak models (decision trees) into a strong predictive model, making it efficient for handling large datasets, missing values, and complex relationships between features. By combining multiple weak models (decision trees) into a single strong predictive model, XGBoost achieves high accuracy and efficiency [8].

2.2 Related Works

Many studies have investigated using machine learning to predict CLV. Traditional methods, like cohort analysis and regression models, are still popular. But they have limitations. These methods struggle with large datasets and complex non-linear relationships. Machine learning approaches, especially ensemble methods, have performed much better at uncovering detailed patterns in customer behavior [9].

For example, Malthouse et al. found that Random Forest and Support Vector Machines worked well in retail and e-commerce [10]. Venkataraman et al. highlighted the impact of gradient boosting in predicting customer churn and retention [11]. Cheng et al. focused on XGBoost for CLV modeling. They pointed out its ability to handle missing data and detect non-linear trends effectively [12]. Still, XGBoost's full potential for CLV prediction remains untapped. This is especially true for customer segmentation and real-world use cases. This research aims to fill those gaps. It shows how XGBoost can be both practical and powerful in predictive analytics.

2.3 Comparative Analysis and Summary

When comparing predictive models for CLV, XGBoost, Gradient Boosting, and K-Nearest Neighbors (K-NN) each have their strengths:

- **Gradient Boosting:** Gradient Boosting is a machine learning technique that iteratively builds models to correct the errors of previous models. This iterative process reduces prediction errors and increases the overall accuracy of the model [13]. However, it can be computationally intensive and is prone to overfitting if not carefully tuned. In this study, Gradient Boosting provided strong results but showed slightly higher misclassification rates compared to XGBoost, particularly for smaller segments.
- **K-Nearest Neighbors (K-NN):** K-NN is a simple and intuitive machine learning algorithm used for classification and regression. It works by finding the 'k' closest data points to a given input and making predictions based on the majority class or average value of those neighbors. K-NN is often preferred for its ease of use and effectiveness, especially when the relationships in the data are complex but not easily modeled with other techniques [14].
- **XGBoost:** XGBoost is a more optimized version of Gradient Boosting, offering improved speed, performance, and accuracy. It can handle missing data, automatically selects features, and is highly scalable for large datasets. Its interpretability, through feature importance analysis, also makes it a valuable tool for understanding customer behavior.

In this study, XGBoost is selected as the primary model due to its superior performance across key metrics like accuracy, precision, and recall.

2.4 Scope of the Problem

This study aims to develop a predictive model for CLV using XGBoost. The approach is tailored for industries with large customer bases, such as e-commerce, telecom, and financial services. While this research primarily addresses digital businesses, its methodology can be adapted to other sectors. Key objectives include:

- Predicting CLV using historical customer data.
- Evaluating the effectiveness of XGBoost in modeling works customer behavior.
- Segmenting customers into groups based on their predicted CLV.

2.5 Challenges

Several challenges arise in implementing this approach:

- **Data Quality:** Missing or poor-quality customer data might lower the model's accuracy and reliability.
- **Feature Selection:** Identifying relevant customer features is the key. Too many irrelevant features may cause overfitting, while too few might miss important details about customer behavior.
- **Model Interpretability:** XGBoost can be complex, making its results harder to understand. Good visualization and explanation tools are needed to address this.
- **Computational Complexity**: Training XGBoost models on big datasets can be resource intensive. Optimizing for both accuracy and efficiency is necessary.
- **Behavioral Variability:** Customer behavior often changes due to external factors. This makes it tricky to create a model that stays accurate over time.

CHAPTER 3
RESEARCH METHODOLOGY

3.1 Research Design

This chapter explains the methodology used to develop a predictive model for Customer Lifetime Value (CLV) and segment customers effectively using machine learning techniques. The main goal is to understand customer behavior and group them into actionable segments. These segments are based on lifetime value, regional preferences, and engagement duration.

The study uses Python-based machine learning tools, clustering methods like K-Means, and classification models like XGBoost. These are applied to gain meaningful insights and make accurate predictions. The workflow includes:

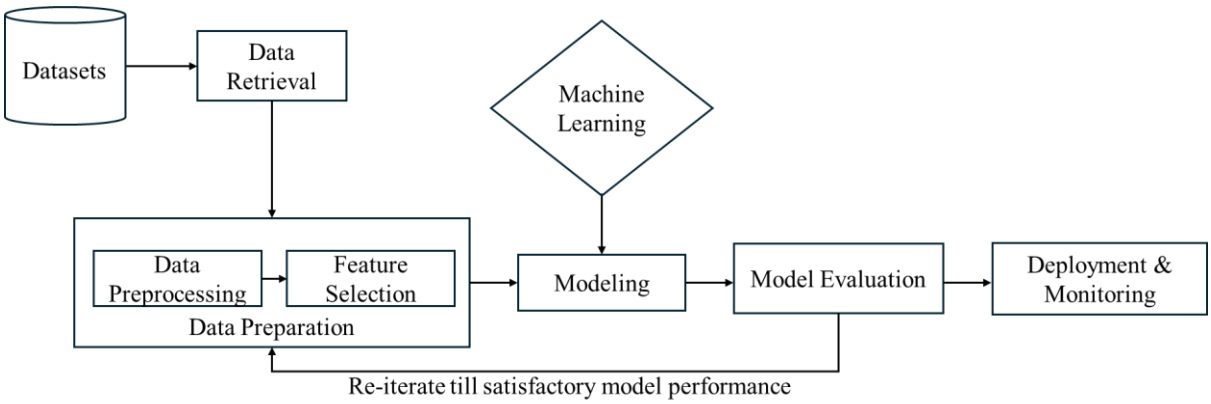

Figure 3.1.1: Research Design

This research uses the following tools and techniques:

- **XGBoost**: The main machine learning algorithm for building the CLV prediction model.
- **K-Means**: Used for clustering customers into distinct groups.
- **Python**: Handles data preprocessing, feature engineering, model training, and evaluation.
- **Jupyter Lab**: Provides an interactive environment for development and visualizations.
- **Libraries**: Key tools include pandas for data handling, scikit-learn for preprocessing and evaluation, and matplotlib and seaborn for creating visualizations.

3.2 Data Collection

The dataset for this research comes from a proprietary source and contains anonymized customer data. It includes key attributes such as Lifetime Value (LTV), Route Region, and customer start and cancel dates.

In table 3.2.1, it demonstrates data set used for K means clustering and XGBoost modeling. There are Unique Customer ID for each customer, the lifetime value they generated, the region they are coming from and their activation and cancellation date.

Customer ID	Lifetime Value (LTV)	Route Region	CustomerStartDate	CustomerCancelledDate
19775	26916	West Lake Hills	10/23/2020	1/1/2090
19201	23929	Rice Military / Heights	2/7/2020	1/1/2090
21991	23865	Galleria	4/21/2021	1/1/2090
20843	21307	Rice Military / Heights	10/8/2020	1/1/2090
30438	20879	Deer Park / La Porte	8/31/2022	1/1/2090
19347	19449	Rice Military / Heights	3/10/2020	1/1/2090
21076	19340	West Lake Hills	11/13/2020	1/1/2090
21783	18399	North Heights / Garden Oaks	2/25/2021	1/26/2024
19127	17336	West U	3/2/2020	11/7/2024
20102	15787	Bear Creek/ Dripping Springs	7/1/2020	1/1/2090
27068	15529	Rice Military / Heights	3/29/2022	1/1/2090
23432	14705	Galleria	7/8/2021	1/1/2090
22830	14473	Galleria	6/2/2021	1/1/2090
20354	12986	North Heights / Garden Oaks	9/23/2020	1/1/2090
20174	12411	Rice Military / Heights	7/14/2020	1/1/2090
20337	12151		7/31/2020	2/9/2023
26716	11997	Galleria	3/4/2022	1/1/2090
21792	11746	Galveston	2/26/2021	1/1/2090
23334	11060	Galleria	7/28/2021	1/1/2090
20046	10907	Rice Military / Heights	7/15/2020	1/1/2090

Table 3.2.1: Preprocessed Data

3.3 Data Preprocessing

Preprocessing ensures that the dataset is clean, consistent, and suitable for analysis. Key statistical techniques include:

- Missing and Invalid cancel dates (e.g., "1/1/2090") were replaced with the current date to ensure accurate age calculations.

- Customer Age was calculated by finding the difference between the customer's start and cancel dates, expressed in years. For example, a customer with a start date of '01/01/2020' and a cancel date of '01/01/2023' would be at an age of 3 years.
- Encoded Route Region (a categorical variable) into numerical values using Label Encoding.
- Numerical features like LTV and Customer Age were scaled using StandardScaler to normalize their distributions, ensuring that all features contribute equally during model training.

In table 3.3.1, it demonstrates the region encoding used in the dataset for K-Means clustering and XGBoost modeling. Each region is assigned a unique numerical code, providing a mapping between region names and their respective encoded values.

Region Name	Region Encoded	Region Name	Region Encoded
Acres Home	0	Humble	25
Addicks / Park Ten	1	Katy	26
Alief	2	Kingwood	27
Bear Creek/ Dripping Springs	3	Lake Conroe/ Willis	28
Bee Cave	4	Lakeway	29
Bellaire	5	Magnolia/ Rose Hill	30
Bluff Springs	6	Manchaca	31
Buda	7	Manor	32
Cedar Park /Leander	8	Memorial	33
Cinco Ranch	9	Mission Bend	34
Conroe	10	Missouri City/Sugarland	35
Cypress North	11	Montrose / South Central Houston	36
Cypress South	12	North Central Austin	37
Deer Park / La Porte	13	North Heights / Garden Oaks	38
Downtown / EADO	14	Northeast Houston	39
Driftwood/ Wimberley	15	Pasadena	40
East Houston / Baytown	16	Pearland	41
Energy Corridor	17	Pecan Grove	42
Friendswood / Dickinson / League City	18	Pflugerville	43
Fulshear	19	Rice Military / Heights	44
Galleria	20	River Oaks	45
Galveston	21	Round Rock	46
Garden Oaks	22	South Central Austin	47
Georgetown	23	South Houston/ Pasadena	48
Hornsby Bend/ Del Valle/ Bastrop	24	Spring Branch	49
Humble	25	Spring/ Louetta	50
		Texas City / La Marque / Bayou Vista	51
		Wells Branch	52
		West Lake Hills	53
		West U	54
		Woodlands/ Tomball	55
			56

Table 3.3.1: Encoded Route Region of the Customers

3.4 Feature Selection

Features used for clustering:

- Lifetime Value (LTV): Directly measures the revenue contribution of each customer.
- Customer Age: Reflects engagement duration, which often correlates with loyalty and spending behavior.
- Route Region (Encoded): Captures geographic trends influencing customer behavior.

3.5 Machine Learning Algorithm

Two machine learning techniques were employed:

1. K-Means Clustering: The dataset was clustered into four groups using the K-Means algorithm, with the number of clusters (k=4) determined through the elbow method, which helps find the optimal k value by evaluating the sum of squared distances. The clustering was based on three features: LTV, Customer Age, and Route Region (rounded). The cluster centers were then analyzed to uncover meaningful patterns in customer behavior.

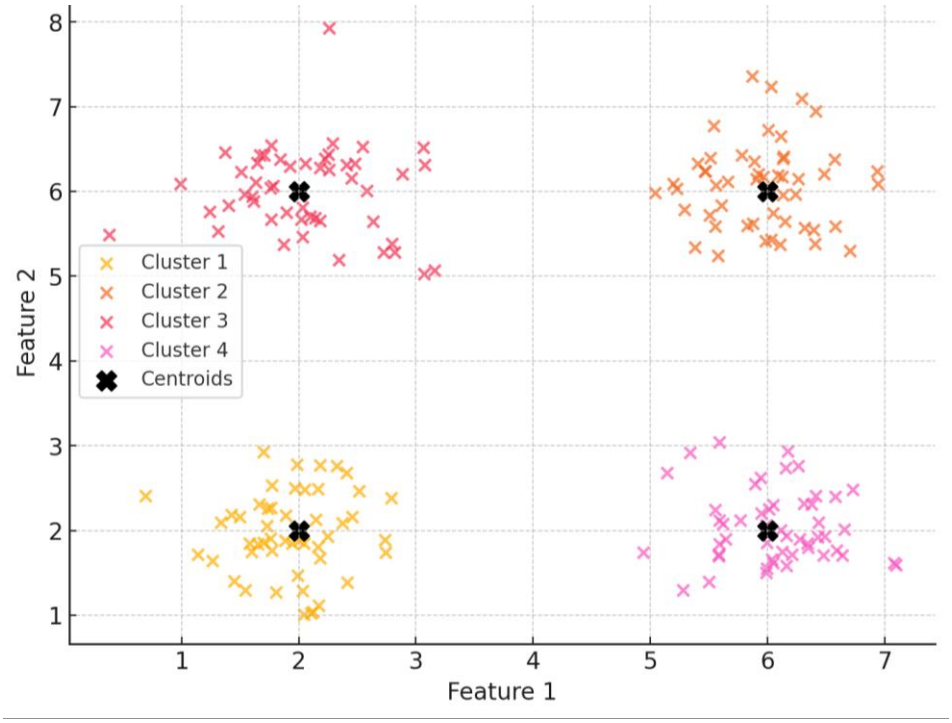

Figure 3.5.1: Visualization of K-Means Clustering

2. XGBoost Classification: The XGBoost classifier was selected for its ability to handle missing data, identify non-linear patterns, and perform well with high-dimensional datasets. It was trained on the scaled feature set to predict cluster membership. The model's performance was evaluated using metrics like Accuracy, Precision, Recall, and F1-Score to ensure robustness and reliability.

3.6 Model Development

Features chosen during the clustering process were used as inputs. K-Means was used for unsupervised clustering, grouping customers into segments based on their LTV, Customer Age, and Route Region. XGBoost, a supervised learning algorithm, was then trained in these segments to classify new customers and predict their CLV.

Flow Chart of the Process:

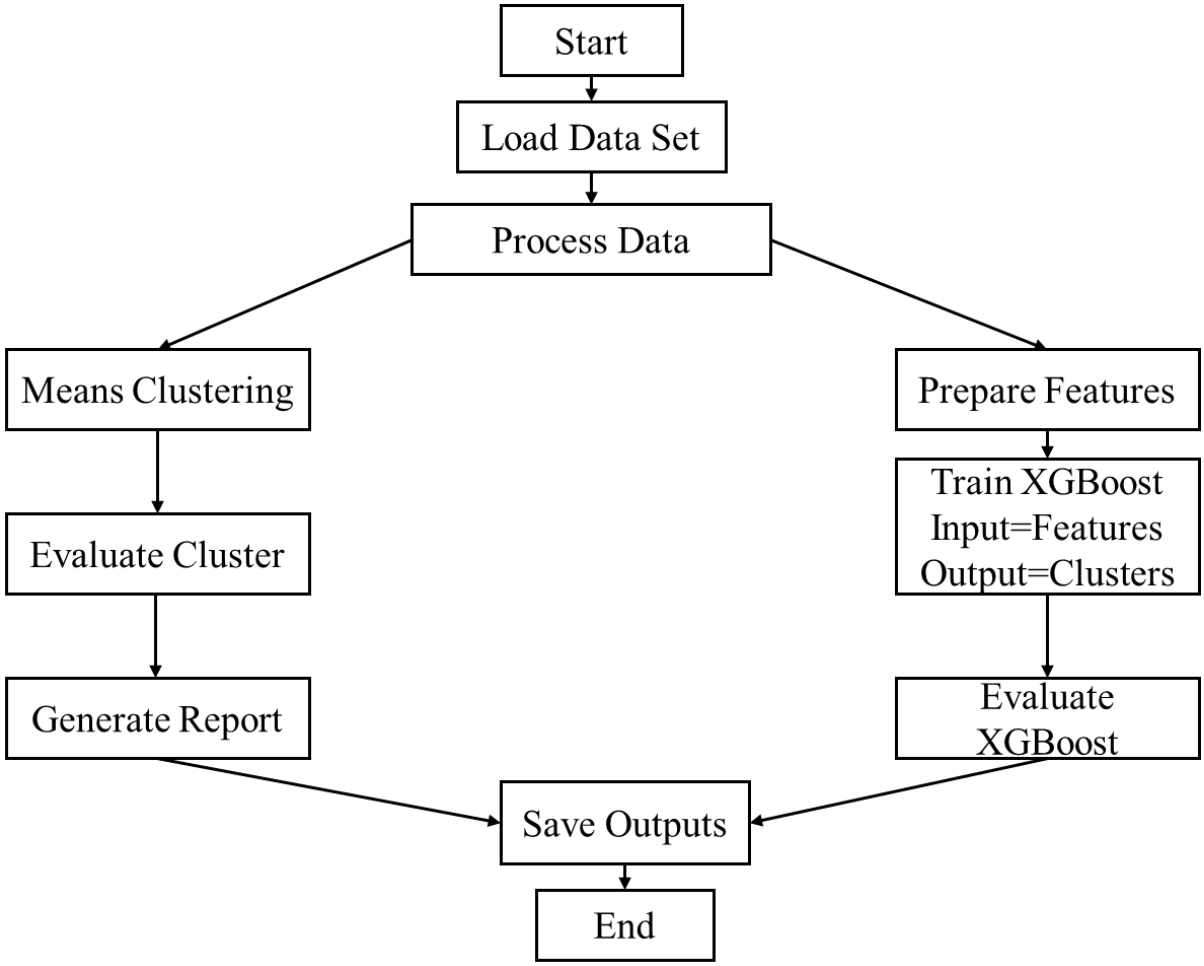

Figure 3.6.1: Model Development & Evaluation

3.7 Classification and Model Evaluation

For this study, three models were tested—XGBoost, Gradient Boosting, and K-Nearest Neighbors (K-NN) to identify the best approach for customer segmentation. Confusion matrices were used to evaluate the performance of each model by analyzing where misclassifications occurred. For example, the confusion matrix for XGBoost showed most predictions aligning with the true segments, as indicated by high values along the diagonal. For example, Segment 2 (high-value customers) had no misclassifications, demonstrating the model's ability to correctly identify smaller but critical groups. To evaluate the performance of each model, key metrics such as Accuracy, Precision, Recall, and F1-Score were used. Accuracy measures the percentage of correctly classified samples, while precision focuses on how many of the predicted positives were correct. Recall evaluates how well the model identifies all true positives, and F1-Score balances precision and recall for a single metric. Together, these metrics provide a comprehensive view of model performance.

CHAPTER 4
EXPERIMENTAL RESULTS AND DISCUSSION

4.1 Dataset Overview

The data set used in this study was obtained from a proprietary source and consists of anonymized customer information. Key features include:

- **Customer Lifetime Value (CLV):** The total revenue generated by each customer.
- **Route Region:** A categorical variable representing the geographical region of customers.
- **Start Date and Cancel Date:** Used to compute the duration of customer engagement (Customer Age).

The dataset underwent preprocessing steps to ensure it was suitable for analysis:

1. Missing values and invalid dates were handled by replacing placeholders with the current date. For example, invalid cancel dates, such as placeholders like '1/1/2090,' were replaced with the current date to ensure accuracy.
2. Categorical features, such as Route Region, were encoded into numerical values using Label Encoding.
3. Numerical features (CLV, Customer Age) were standardized using StandardScaler for consistency,

4.2 Performance Metrics

The evaluation of clustering and classification models relied on the following performance metrics:

1. **Root Mean Square Error (RMSE):** Measures how well the clustering algorithm grouped the data. A lower value indicates more compact clusters. The RMSE for K-Means was 0.5917, indicating compact and well-defined clusters.
2. **Accuracy:** Represents the proportion of correctly classified samples.
3. **Precision:** Measures the proportion of true positive predictions among all positive predictions for each segment.
4. **Recall:** Measures the ability of the model to correctly identify all positive samples within a segment.

5. **F1-Score:** Balances precision and recall into a single metric.

4.3 Interpretation of Findings

4.3.1 Clustering Insights:

The following table 4.3.1 summarizes the characteristics of the four customer segments:

Segment	Mean LTV	Median LTV	Std. LTV	Count	Mean Customer Age	Most Common Region	Region Count
0	2284.91	2143	980.27	1863	3.19	38	180
1	693.65	593	536.15	2515	0.97	56	459
2	7435.71	6440	3351.78	247	3.43	20	33
3	810.77	679	628.54	1885	1.06	5	232

Table 4.3.1: Characteristics of the four customer segments.

4.3.2 Interpretation of Clusters:

Segment 0: This segment includes customers with moderate lifetime value and engagement duration. Region 38 is the most common for customers in this group, with 180 customers.

Segment 1: These are younger customers with low lifetime value. This is the largest segment, comprising 2515 customers, predominantly from region 56.

Segment 2: This group contains high-value customers with longer engagement periods. Although this segment is small (247 customers). Region 20 is the most common. This segment includes the most valuable customers. Businesses should focus on loyalty programs or exclusive offers to retain these individuals.

Segment 3: Slightly older customers with relatively low lifetime value. Most of these customers are from region 5, totaling 232.

The following Table 4.3.2 presents the output of K-Means segmentation for each customer. A new column has been added to the dataset to indicate the segment number, representing the cluster to which each customer belongs.

Customer ID	Lifetime Value (LTV)	Route Region	CustomerStartDate	CustomerCancelledDate	CustomerAge	RouteRegionEncoded	Segment
386	7486	Bellaire	2020-03-06	2024-12-15 01:06:52.788541	4.777549623545520	5	2
1027	1386	West U	2022-05-21	2024-12-15 01:06:52.788541	2.5708418891170400	54	0
1270	1292	West U	2024-01-29	2024-12-15 01:06:52.788541	0.8788501026694050	54	1
1334	1133	Woodlands/ Tomball	2023-04-29	2024-12-15 01:06:52.788541	1.6317590691307300	55	1
2179	628	Cypress North	2023-09-06	2024-12-15 01:06:52.788541	1.2758384668035600	11	3
2922	799	Rice Military / Heights	2021-09-27	2022-12-20 00:00:00.000000	1.2292950034223100	44	1
3295	6355	Woodlands/ Tomball	2021-06-11	2024-11-08 00:00:00.000000	3.4113620807666000	55	2
3462	57	River Oaks	2024-11-19	2024-12-15 01:06:52.788541	0.07118412046543460	45	1
3749	1629	Rice Military / Heights	2022-10-03	2024-12-15 01:06:52.788541	2.2012320328542100	44	0
3961	2574	River Oaks	2021-10-26	2024-12-15 01:06:52.788541	3.137577002053390	45	0
3976	1710	North Heights / Garden Oaks	2021-09-16	2024-08-04 00:00:00.000000	2.8829568788501000	38	0
4706	2526	Bellaire	2021-03-31	2024-12-15 01:06:52.788541	3.7097878165640000	5	0
4716	1681	Friendswood / Dickinson / League City	2020-07-07	2024-12-15 01:06:52.788541	4.440793976728270	18	0
4766	1948	North Heights / Garden Oaks	2020-05-01	2022-04-14 00:00:00.000000	1.9520876112251900	38	0
4784	2724	Rice Military / Heights	2020-09-24	2024-12-15 01:06:52.788541	4.224503764544830	44	0
4848	232	Garden Oaks	2021-10-14	2022-03-02 00:00:00.000000	0.3805612594113620	22	3
4964	1711	West U	2020-09-03	2024-12-15 01:06:52.788541	4.2819986310746100	54	0

Table 4.3.2: Segment for each customer

Feature Pairwise Distribution (Pairplot/Scatter Matrix):

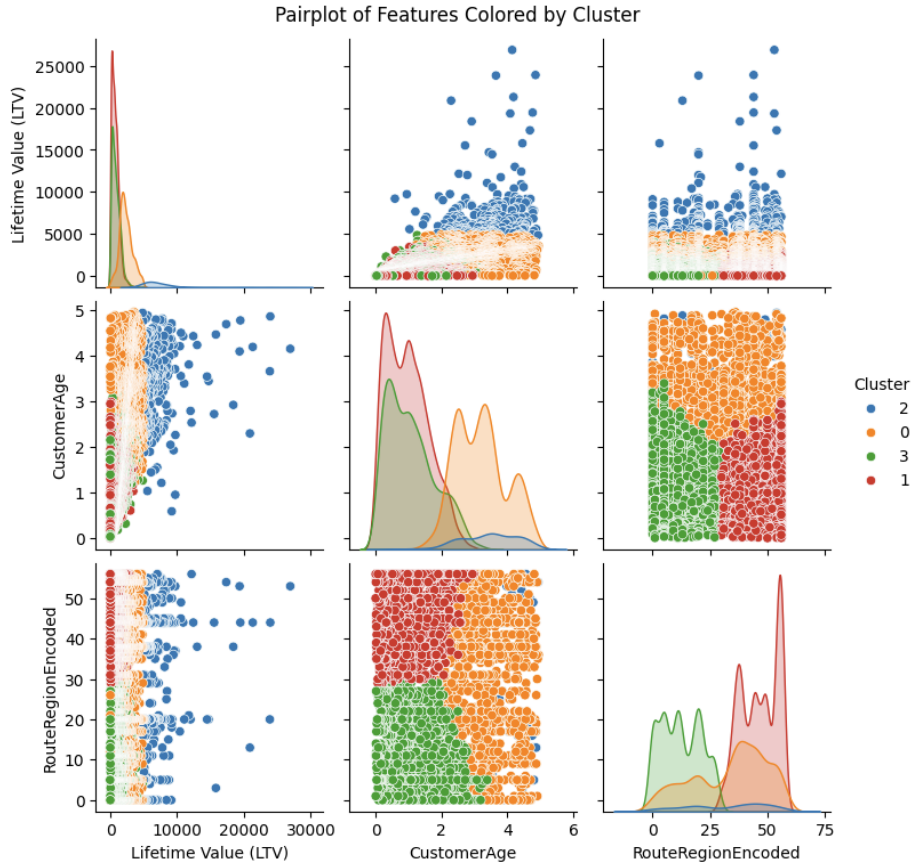

Figure 4.3.2: Feature Pairwise Distribution

Diagonal Plots (Feature Distributions):

Lifetime Value (LTV):

- Cluster 2 (blue) includes high-value customers with a long tail, suggesting a premium or high-spending segment.
- Cluster 3 (green), 1 (red), and 0 (orange) are more concentrated at lower LTV values.

Customer Age:

- Cluster 2 (blue) corresponds to older customers.
- Cluster 3 (green) represents relatively younger customers, with Clusters 0 (orange) and 1 (red) distributed across medium and younger age ranges.

Route Region Encoded:

- Different clusters dominate specific ranges of encoded regions, suggesting a strong regional factor in segmentation.

Off-Diagonal Relationships:

Lifetime Value (LTV) vs. Customer Age:

- Positive correlation observed for Cluster 2 (blue), where older customers tend to have higher LTV.
- Other clusters show scattered, lower LTV values, with little correlation to age.

Lifetime Value (LTV) vs. Route Region Encoded:

- Clusters 2 and 0 have spread across various regions, while Clusters 1 and 3 dominate specific regions, suggesting regional preferences or customer behavior.

Customer Age vs. Route Region Encoded:

- Clusters segregate well by region, with distinct age groups visible. Cluster 2 includes older customers across diverse regions, while Clusters 0, 1, and 3 show localized age and regional patterns.

Cluster-Specific Observations:

- **Cluster 2 (Blue)**: High-value, older customers spread across multiple regions. Likely the most valuable segment.
- **Cluster 0 (Orange):** Mid-range value and age, with concentration in specific regions.
- **Cluster 1 (Red):** Likely young and low-value customers confined to a limited region.
- **Cluster 3 (Green):** Newer or younger customers with medium value, dominating certain regions.

1. **XGBoost classifier's performance:**

The performance of the XGBoost classifier was evaluated using precision, recall, F1-score, and overall accuracy. In Table 4.3.3, the results are summarized below:

Metric	Segment 0	Segment 1	Segment 2	Segment 3	Overall
Precision	0.97	0.99	1.00	1.00	99%
Recall	0.99	0.99	0.98	0.98	99%
F1-Score	0.98	0.99	0.99	0.99	99%
Support (Count)	377	494	60	371	1302

Table 4.3.3: XGBoost classifier's performance

Key Observations:

- XGBoost demonstrated the highest overall accuracy (98.62%), with minimal misclassifications.
- Smaller yet critical segments, like Segment 2, were classified with near-perfect precision and recall.
- Gradient Boosting performed comparably but had higher misclassification rates, especially for Segment 0 and Segment 3.

Performance Analysis Based on Confusion Matrices and Classifier Metrics:

1. XGBoost

Confusion Matrix Analysis: Very few misclassifications

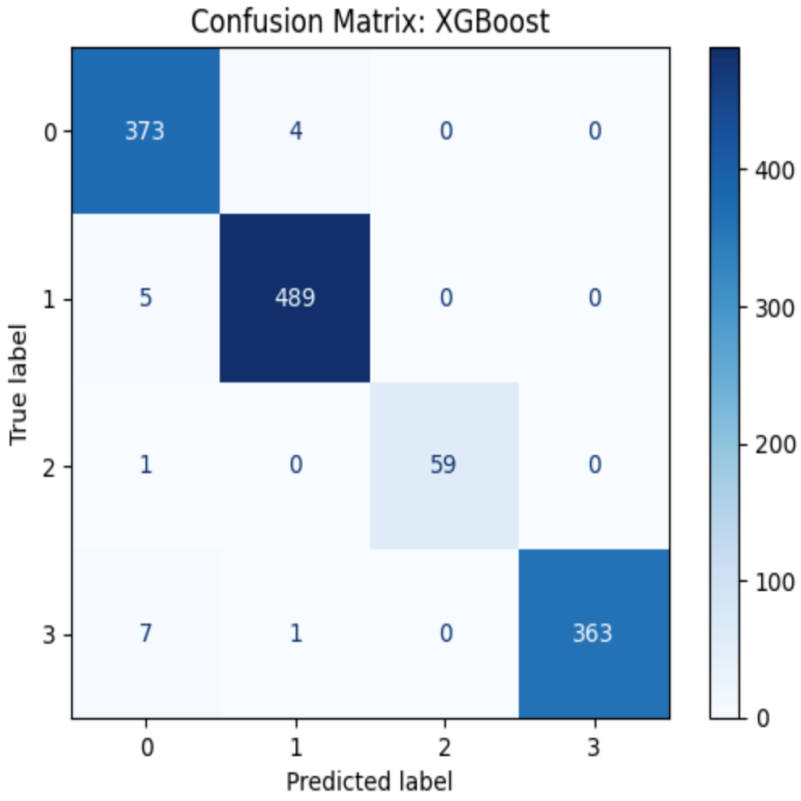

Figure 4.3.3: Confusion matrix of XGBoost

Segment 0 had 4 instances misclassified as Segment 1, while Segment 1 had 5 instances misclassified as Segment 0. Segment 3 showed 7 instances misclassified as Segment 0. There were only 1 misclassification for Segment 2, demonstrating excellent precision and recall for smaller segments. Overall, the confusion matrix showed diagonal dominance, indicating that most predictions aligned correctly with the true classes.

Classifier Metrics: The overall accuracy of the model is 98.62%. Precision ranges from 97% to 100% across clusters, while recall falls between 98% and 99%. The F1-Score is consistently 99% for all clusters. XGBoost provides outstanding classification performance, handling all segments effectively, including smaller ones like Segment 2. In table 4.3.4, the results are summarized below:

XGBoost Performance			
Segment	Precision	Recall	F1-Score
0	0.97	0.99	0.98
1	0.99	0.99	0.99
2	1.00	0.98	0.99
3	1.00	0.98	0.99

Table 4.3.4: XGBoost Performance

The overall accuracy of the model is 98.62%. Precision ranges from 97% to 100% across clusters, while recall falls between 98% and 99%. The F1-Score is consistently 99% for all clusters. XGBoost provides outstanding classification performance, handling all segments effectively, including smaller ones like Segment 2.

2. Gradient Boosting

Confusion Matrix Analysis: Misclassification rates are slightly higher than XGBoost.

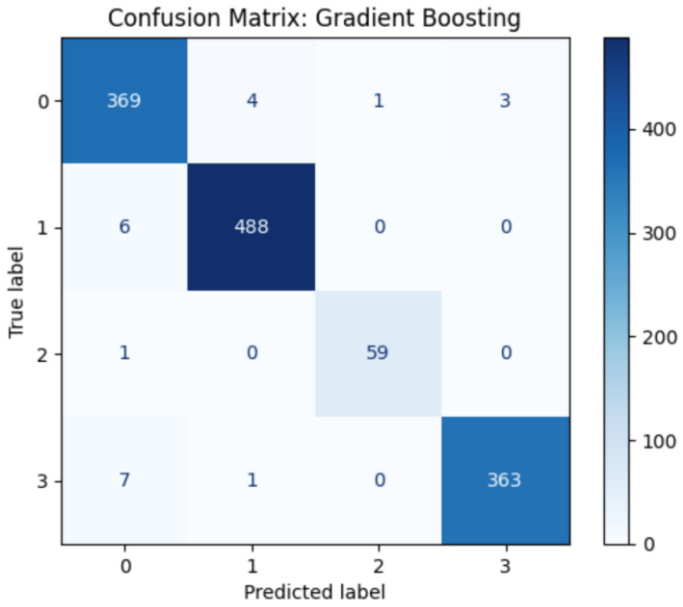

Figure 4.3.4: Confusion matrix of Gradient Boosting

Segment 0 had 4 instances misclassified as Segment 1, 1 as Segment 2, and 3 as Segment 3. Segment 1 showed 6 instances misclassified as Segment 0, while Segment 3 had 7 instances

misclassified as Segment 0. Segment 2 performed strongly with only 1 misclassification. The confusion matrix still showed diagonal dominance, though there were slightly more errors compared to XGBoost.

Classifier Metrics: Overall accuracy is 98.23%. The macro average precision, recall, and F1-score are 0.98 across all segments. Class-wise performance is slightly lower than XGBoost, particularly for Segment 0 and Segment 3. Overall, Gradient Boosting is close second to XGBoost, offering comparable metrics but with slightly higher misclassification rates. In table 4.3.5, the results are summarized below:

Gradient Boosting Performance			
Segment	Precision	Recall	F1-Score
0	0.96	0.98	0.97
1	0.99	0.99	0.99
2	0.98	0.98	0.98
3	0.99	0.98	0.99

Table 4.3.5: Gradient Boosting Performance

3. K-Nearest Neighbors

Confusion Matrix Analysis: Significant misclassifications

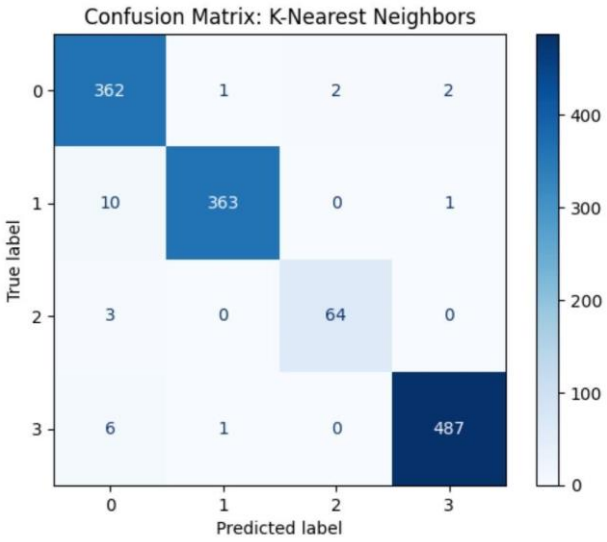

Figure 4.3.5: Confusion matrix of K-Nearest Neighbors

Segment 0 had 1 instance misclassified as Segment 1, 2 as Segment 2, and 2 as Segment 3. Segment 1 showed 10 misclassifications as Segment 0 and 1 as Segment 3. Segment 2 had 3 instances misclassified as Segment 0. Segment 3 showed 6 misclassifications such as Segment 0 and 1 as Segment 1. K-NN performed well overall but exhibited slightly more misclassifications compared to Gradient Boosting and XGBoost.

Classifier Metrics: Overall accuracy is 98%. The macro average precision is 0.98, recall is 0.97, and F1-score is 0.98. Segment 3 performed the best with precision, recall, and F1-score all at 0.99. Segment 0 had a slightly lower precision at 0.95 but a high recall at 0.99. Segment 1 showed a strong performance with precision at 0.99 and recall at 0.97. Segment 2 had slightly lower scores, with precision at 0.97 and recall at 0.96. K-NN performed well overall, but had slightly more misclassifications compared to XGBoost, especially in smaller segments. In table 4.3.6, the results are summarized below:

K-Nearest Neighbors			
Segment	Precision	Recall	F1-Score
0	0.95	0.99	0.97
1	0.99	0.97	0.98
2	0.97	0.96	0.96
3	0.99	0.99	0.99

Table 4.3.6: K-Nearest Neighbors Performance

Comparison of models: In table 4.3.7, it demonstrates the comparison of XGBoost performance with Gradient Boosting and K-Nearest Neighbors Model:

Model	Accuracy	Precision (Macro Average)	Recall (Macro Average)	F1-Score (Macro Average)
XGBoost	0.9862	0.99	0.99	0.99
Gradient Boosting	0.9823	0.98	0.98	0.98
K-Nearest Neighbors	0.9800	0.98	0.97	0.98

Table 4.3.7: Model Results Overview

Summary: XGBoost outperformed Gradient Boosting and K-NN with an overall accuracy of 98.62% and near-perfect Precision, Recall, and F1-Scores across all segments. It handled small, high-value segments like Segment 2 with ease, while K-NN struggled with misclassifications due to its sensitivity to noisy data. Confusion matrices provide a detailed view of how well a model predicts each class. For example, in the XGBoost confusion matrix, most predictions align with true classes (seen along the diagonal), indicating high accuracy. However, K-NN had more off-diagonal entries, meaning it struggled to classify certain segments, such as Segment 0 and Segment 3. Overall, XGBoost stands out as the most effective model for this study.

Recommendation: Based on the evaluation results, XGBoost is recommended as the most effective model for customer segmentation. It demonstrated the highest accuracy and consistently outperformed Gradient Boosting and K-NN across all key metrics, including precision, recall, and F1-score. XGBoost's ability to handle complex data patterns and minimize misclassifications makes it a reliable and efficient choice. Therefore, it should be implemented for tasks requiring accurate customer segmentation and prediction.

4.4 Discussion of Findings

Using K-Means clustering and XGBoost classification together gave clear insights and reliable predictions for customer segmentation. The main takeaways are:

- Segment 1 (largest group) offers growth opportunities through targeted marketing strategies.
- Segment 2 (high-value customers) should be the primary focus for retention programs to maximize long-term profitability.

4.4.1 Implications for Findings:

- Segment 0 and Segment 1 cover most customers. They have moderate to low lifetime value, so focusing on retention and upselling could help.
- Segment 2, comprising high-value customers with the longest engagement periods, represents a critical group for businesses. Targeting this segment with personalized loyalty programs and exclusive offers can maximize retention and long-term profitability.
- Segment 3 are slightly older customers and have lower lifetime value. Re-engagement efforts may be effective for them.

CHAPTER 5
IMPACT ON SOCIETY, ENVIRONMENT AND SUSTAINABILITY

5.1 Impact on Society

The predictive Customer Lifetime Value (CLV) models developed in this study bring positive changes to society by helping businesses better understand and serve their customers. By using machine learning methods like XGBoost and K-Means, companies can provide more personalized and efficient services, benefiting both customers and businesses.

- **Better Customer Experiences**: Improved segmentation allows businesses to offer services, products, or promotions that align with customer needs. This personalized approach builds trust and increases satisfaction.
- **Community Engagement**: Companies can identify underserved groups or niche markets. This leads to more inclusive products and services, fostering stronger connections with communities.
- **Economic Growth**: Accurate predictions help businesses optimize resources, boost profits, and expand. This growth can lead to job creation and other economic benefits.
- **Support for Small Businesses**: These models help small businesses compete with larger companies. By identifying high-value customers, they can focus on retention and growth, making data-driven strategies accessible to all.

Still, challenges like protecting consumer privacy and addressing digital inequality must be tackled. This ensures that the benefits are shared fairly.

5.2 Impact on Environment

This study mainly looks at customer segmentation and CLV prediction, but the environmental impact of these models should also be considered. Potential Benefits:

- **Energy Efficiency:** Models like XGBoost and K-Means can help businesses save energy. With accurate predictions, supply chains can be optimized, and inventory management can improve. For example, better demand forecasts reduce overproduction, which lowers waste and cuts carbon emissions.
- **Reduced Carbon Footprints:** Focusing on the right customers can reduce mass advertising. This means less energy use from data centers and servers.
- **Support for Sustainability:** Insights from segmentation can push businesses toward sustainability. High-value customers, like those in Segment 2, might prefer eco-friendly products. This could lead companies to create more sustainable goods or invest in carbon-offset

programs. Predictive models can identify which customers are more likely to support green products.

Training complex machine learning models takes a lot of energy. High-performance computing increases carbon emissions. To manage this, companies can use renewable energy, make training processes more efficient, and adopt energy-saving infrastructure.

5.3 Ethical Aspects

Using machine learning for predicting Customer Lifetime Value (CLV) brings up some ethical issues. Businesses need to be mindful of these:

- **Data Privacy and Security:** Customer data is essential for this study, but it must be handled carefully. Following rules like GDPR and CCPA is important to protect personal data. Companies should be transparent to avoid misuse.
- **Bias and Fairness:** Machine learning models like XGBoost rely on the data they are trained with. If the data is biased, the results could favor some groups, like certain regions or age groups, while ignoring others. To prevent this, businesses need to check for bias and adjust the data to ensure fairness.
- **Customer Autonomy and Consent:** Predictive models often use customer data for personalized services. Customers should have control over how their data is used. Businesses must get clear consent and let customers opt-out or delete their data without facing any issues.
- **Transparency in Model Decisions:** Advanced models can feel confusing, like a "black box." People may not understand how decisions are made. This becomes a problem if errors happen. Companies should make these models easier to explain so customers can understand and challenge decisions when needed.

5.4 Sustainability Plan

To align this study's findings with long-term sustainability goals, businesses need a structured approach. Both environmental and social factors should be considered. Here are some key actions:

- **Optimizing Energy Use in Machine Learning:** Companies can cut energy use by using renewable-powered cloud platforms and optimizing models to reduce training time. Companies can cut energy use by using renewable-powered cloud platforms and optimizing models to reduce training time.
- **Sustainable Customer Engagement:** Promote sustainable customer engagement by focusing on high-value clients (Segment 2) through eco-friendly products and services. Provide

incentives for sustainable practices and inform customers about the environmental consequences of their choices.

- **Inclusive Practices:** Equity in consumer segmentation is essential. All demographic groupings ought to be represented equitably. Concentrate on meeting the requirements of marginalized groups and guarantee that products and marketing are universally accessible.
- **Continuous Monitoring and Improvement:** Consumer preferences and technology evolve over time. Businesses need to periodically refresh predictive models to maintain their relevance and sustainability. Consistent assessments facilitate the achievement of contemporary objectives.
- **Collaboration with Stakeholders:** Sustainability necessitates collective effort. Businesses must collaborate with governmental entities, non-profit organizations, and industry associations to guarantee that their models are advantageous to society and the environment. Collaborative efforts can also establish ethical principles for machine learning.

CHAPTER 6
SUMMARY AND FUTURE WORK

6.1 Summary of the Study

This research focused on creating a predictive Customer Lifetime Value (CLV) model using K-Means clustering and XGBoost classification. The primary objective was to improve customer segmentation and provide actionable insights for businesses. The methodology encompassed data preprocessing, which included managing categorical variables, scaling, and feature selection. Customers were categorized into four groups using K-Means, depending on parameters such as Lifetime Value (LTV), Customer Age, and Route Region. These segment labels were then used as targets for the XGBoost classifier, which demonstrated exceptional accuracy in forecasting client segments. The performance of the clustering model was evaluated using Root Mean Square Error (RMSE), yielding a value of 0.5917, which signifies well-defined clusters. The XGBoost classifier attained 99% accuracy, exhibiting robust precision, recall, and F1-scores, thus demonstrating its efficacy in customer classification and future behavior prediction. The analysis highlighted interesting customer group trends. Segment 2 included high-value customers with the longest engagement and highest Lifetime Value, making them a key focus for retention. Segment 1, the largest group, consists of younger customers with lower Lifetime Value, offering opportunities for targeted growth strategies. Beyond these findings, the study highlighted broader impacts. Accurate segmentation and CLV predictions can improve customer satisfaction, minimize waste, and encourage sustainable business practices.

6.2 Conclusions

This study shows that using K-Means clustering with XGBoost offers a strong method for customer segmentation and revenue forecasting. The main findings are:

K-Means clustering effectively divided customers into four meaningful segments based on their behavioral and demographic attributes. Segment 2 stood out for its high-value potential, while Segment 1 represented the largest, lower-value group, emphasizing diverse strategic needs for each segment. The XGBoost classifier demonstrated exceptional accuracy in predicting customer segments, attaining an overall accuracy of 99%. This underscores the ability of sophisticated machine learning models to facilitate data-driven decision-making in

customer relationship management. The segmentation model offers explicit direction for marketing strategies and customer retention efforts. Lower-value groups may be addressed through re-engagement or retention initiatives, whilst high-value segments could gain from tailored loyalty programs. The study highlights the capacity of these models to enhance operational efficiency and promote sustainable customer interaction, enabling firms to make a beneficial impact on society and the environment.

6.3 Implications for Further Study

This study demonstrates the effectiveness of the proposed customer segmentation model, but there are still areas that could be explored in future research. Here are a few ideas:

Future research might consider utilizing real-time data, such as online browsing or social media interactions. This might help to reveal more about what customers like and how they decide to buy things. Even though K-Means performed effectively, it might be worth trying out methods such as DBSCAN or hierarchical clustering. These could be more effective at managing complex or non-linear patterns. Long-term studies can help to understand how customer behavior and lifetime value change over time, showing how different segments develop. This would assist companies in modifying their strategies as required. This study concentrated on a single industry; nevertheless, the approach may be applicable in sectors such as retail, banking, healthcare, or telecommunications. Subsequent study may investigate methods for contextual adaptation. XGBoost showed great accuracy; yet machine learning models often lack clarity in interpretation. Future endeavors may concentrate on enhancing the comprehensibility of predictions by employing tools such as SHAP to elucidate the impact of attributes. Research must address ethical considerations, such as mitigating data bias and guaranteeing equitable treatment of all demographic groups. Machine learning models possess a significant ecological impact. Subsequent research may concentrate on enhancing algorithms to conserve energy and minimize computational expenses. Research could also explore how to connect predictive models with tools like CRM or ERP systems. This integration could improve operations and enable real-time customer segmentation.

REFERENCES

[1] Kumar, V., & Shah, D. (2004). *Building and Sustaining Profitable Customer Loyalty for the 21st Century*. Journal of Retailing, 80(4), 317-330.

[2] Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.

[3 Liu, B., & Wang, X. (2018). Predicting Customer Lifetime Value with Deep Learning: A Comparative Study. International Journal of Information Technology & Decision Making, 17(4), 1345-1365.

[4] Malthouse, E. C., & Blattberg, R. C. (2013). Can we predict customer lifetime value? Journal of Interactive Marketing, 27(4), 101-111.

[5] Gupta, S., et al. (2006). Modeling customer lifetime value. *Journal of Marketing Research, 43*(1), 14-31.

[6] Wedel, M., & Kamakura, W. A. (2000). Market segmentation: Conceptual and methodological foundations. *Kluwer Academic Publishers*.

[7] MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*.

[8] Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

[9] Venkataraman, S., et al. (2007). Influencing customer lifetime value through marketing. *Marketing Letters, 18*(3), 189-203.

[10] Malthouse, E. C., & Blattberg, R. C. (2013). Can we predict customer lifetime value? *Journal of Interactive Marketing, 27*(4), 101-111.

[11] Larivière, B., & Van den Poel, D. (2005). Predicting retention with random forests. *Expert Systems with Applications, 29*(2), 472-484.

[12] Cheng, J., et al. (2020). Leveraging machine learning to enhance CLV prediction. *Journal of Data Science and Applications*.

[13] Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics, 29*(5), 1189-1232.

[14] Cover, T. and Hart, P. (1967). *Nearest neighbor pattern classification*. IEEE Transactions on Information Theory, 13(1), 21-27.

Customer Lifetime Value Modeling: A Machine Learning Approach To Customer Segmentation By K-Means And XGBoost

ORIGINALITY REPORT

14%	11%	7%	6%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Submitted to Daffodil International University Student Paper	2%
2	dspace.daffodilvarsity.edu.bd:8080 Internet Source	1%
3	openaccess.altinbas.edu.tr Internet Source	1%
4	www.mdpi.com Internet Source	1%
5	www.researchsquare.com Internet Source	1%
6	Submitted to Virginia Commonwealth University Student Paper	1%
7	Ramcharan Kakarla, Sundar Krishnan, Balaji Dhamodharan, Venkata Gunnu. "Chapter 11 Modeling Frameworks", Springer Science and Business Media LLC, 2024 Publication	1%