ADDIS ABABA UNIVERSITY

COLLEGE OF NATURAL AND COMPUTATIONAL SCIENCES

SCHOOL OF INFORMATION SCIENCE

Applying Data Mining Technology for Customer Lifetime Value (CLV) Prediction: The Case of Commercial Bank of Ethiopia (CBE)

BY: TRUALEM SISAY

ID GSE/2702/13

ADVISOR: MARTHA YIFIRU (Ph.D.)

October 2023

ADDIS ABABA, ETHIOPIA

ADDIS ABABA UNIVERSITY

COLLEGE OF NATURAL AND COMPUTATIONAL SCIENCES

SCHOOL OF INFORMATION SCIENCE

Applying Data Mining Technology for Customer Lifetime Value (CLV) Prediction: The Case of Commercial Bank of Ethiopia (CBE)

A Thesis Submitted to the School of Information Science of Addis Ababa University in Partial Fulfilment of the Requirements for the Degree of Master of Science in Information Science and Systems (Information Systems)

BY: TRUALEM SISAY

October,2023

ADDIS ABABA UNIVERSITY

COLLEGE OF NATURAL AND COMPUTATIONAL SCIENCES

SCHOOL OF INFORMATION SCIENCE

Applying Data Mining Technology for Customer Lifetime Value (CLV) Prediction: The Case of Commercial Bank of Ethiopia (CBE)

BY: TRUALEM SISAY

APPROVAL SHEET

NAME AND SIGNATURE OF MEMBERS OF THE EXAMINING BOARD

_____	_____	_____
Advisor	Signature	Date
_____	_____	_____
Examiner	Signature	Date
_____	_____	_____
Examiner	Signature	Date

Declaration

I, the undersigned, hereby certify that the thesis titled "Applying Data Mining Technology for Customer Lifetime Value (CLV) Prediction: The Case of Commercial Bank of Ethiopia (CBE)" is my own original work. No university has accepted it as one of the requirements for a degree. I properly cited each source of information used in the thesis.

_____

Trualem Sisay October 2023

With my consent as the university advisor, the thesis has been turned in for review.

Name: _____

Signature: _____

Date: _____

Acknowledgment

I am forever grateful to my almighty God for providing everything I needed to complete this thesis. I want to thank everyone for their praise and assistance throughout this journey. I am extremely thankful to my thesis advisor, Dr. Martha Yifiru, for her valuable feedback, advice, and support. Without her guidance and patience, this thesis would not have been possible from start to finish.

I want to sincerely acknowledge the Information Management (IM) staff at the Commercial Bank of Ethiopia for their backing and understanding throughout my journey. I am eternally thankful to my friends and coworkers for constantly inspiring and encouraging me. They have been by my side through both good times and bad, offering steadfast assistance whenever I needed it most. Their confidence in me has always fueled my progress towards reaching my goals.

In conclusion, the extent of my gratitude for my family is beyond words. They have played a big role in positively shaping my life. Their love, kindness, and support mean everything to me!

Table of Contents

List of Figures

List of tables

List of Acronyms and Abbreviations

ATM	Automated Teller Machine
CBE	Commercial Bank of Ethiopia
CLV	Customer Lifetime Value
CPU	Central Processing Unit
CRISP-DM	Cross Industry Standard Process Data Mining
CRM	Customer Relationship Management
DM	Data Mining
DSP	Design Science Process
DSRM	Design Science Research Methodology
ETL	Extract, Transform, and Load
EDW	Enterprise Data Warehouse
EDA	Exploratory Data Analysis
GB	Giga Byte
IM	Information Management
KDD	Knowledge Discovery in Database
MAE	Mean Absolute Error
ML	Machine learning
MSE	Mean Squared Error
ODI	Oracle Data Integrator
R^2	Coefficient of Determination or R squared Value
RAM	Random Access Memory
RFE	Recursive Feature Elimination

RMSE	Root Mean Squared Error
SQL	Structured Query Language
SEMMA	Sample Explore Modify Model Assess

Abstract

Data mining is a powerful tool for businesses to discover hidden patterns and insights from their data. This can be used to improve customer relationship management (CRM), particularly in understanding the key factors that contribute to predicting customer lifetime value (CLV). CLV is the present value of all future profits that a customer will generate throughout their relationship with a company. One of the various service sectors that gathers, manages, and retains enormous volumes of data over time is the Commercial Bank of Ethiopia (CBE). This data can be used to predict CLV and improve CRM by understanding the key factors that influence customer behavior.

This study used data mining techniques to predict CLV at CBE using the Cross-Industry Standard Process for Data Mining (CRISP-DM). The dataset included information on 100,096 customers and 19 attributes, such as demographics, account details, usage, and transactions. After business understanding and data understanding the data was prepared for experimentation. This involved removing outliers and converting categorical variables to numerical values. Based on their benefits and prior performance reported in the literature, the three machine learning algorithms linear regression, random forest, and decision tree were chosen for the experiment. The algorithms were implemented and their performance was assessed using the Python programming language. Using R2 and RMSE, the models' performance was assessed.

The results revealed that random forest regression had the highest R^2, at 86.8%, followed by decision tree regression at 72.4% and linear regression at 56.91%. The study also found that the most important features for CLV prediction are transaction-related features, such as debit amount, credit amount, and total number of transactions. This study demonstrates the applicability of data mining techniques to improve customer relationship management at CBE. By understanding the key factors that contribute to CLV, CBE can develop targeted interventions to keep its customers engaged and grow its business.

Keywords: Data Mining, Machine Learning Algorithms, Customer Lifetime Value

CHAPTER ONE

INTRODUCTION

1.1 Background

Numerous studies have shown that keeping existing customers is less expensive than obtaining new ones. As a result, a key element in determining whether a business succeeds or fails is evaluating current customers in order to retain high-value customers and increase their lifetime value. Customer lifetime value (CLV), which represents a customer's profitability over the course of their engagement with a business, is a key term in the field of customer relationship management (CRM). Customer relationship managers can assess present and projected customer profitability using CLV as a key metric, and they can tailor products and services to attract and keep customers who are profitable [1]. Businesses are switching from using mass marketing to using one-to-one marketing. The main objective of a business under the conventional model is to increase customer interaction and grow its business. The cost of selling to existing customers has, however, forced enterprises to concentrate their efforts on doing so due to growing competition. The deliberate efforts to sustain positive customer relationships, along with cutting-edge techniques and technologies, have given rise to the field of customer relationship management [2].

Data-mining applications are becoming more prevalent in a variety of industries, including retail, banking, telecommunications, and supply chain management. With the help of store-branded credit cards and a point-of-sale system, companies can collect and maintain records of every transaction. Such information has been used for basket analysis, sales forecasting, database marketing, merchandise planning, and other purposes. Knowledge discovery techniques can be employed in domains such as banking, credit card marketing, fraud detection, telecommunications, customer loyalty, churn prediction, and promotion response prediction [2]. Technological advancements have led to the emergence of numerous new fields. Every day, various fields, such as science, engineering, health, and business, generate and accumulate massive amounts of data. Data mining is an important field that manages and extracts necessary information from large amounts of data [3]. Data mining techniques help extract patterns from large datasets and have been applied to predict CLV parameters and resource allocation.

Techniques such as logistic regression, decision trees, random forests, artificial neural networks, genetic algorithms, and support vector machines have been used for this purpose [2].

Data mining is a powerful tool that can be used to address various business challenges. It can be used to create sustainable competitive advantages, such as improved customer relationship management, customer profiling, churn analysis, and identification of the most profitable customers. By predicting CLV, businesses can develop more targeted marketing campaigns that help them retain customers for longer. Additionally, data mining can be used to identify associations between products and services to maximize sales [5]. The banking industry is highly competitive, and banks are constantly looking for ways to increase customer loyalty and maximize CLV. By predicting CLV, banks can identify high-value customers and target them with personalized offers and services. This study aims to apply data mining techniques to CLV prediction for CBE. This will help CBE to identify valuable customers and improve its marketing strategy.

CBE was founded in 1942 and has played an important role in Ethiopia's development. CBE currently has more than 39.9 million account holders in its 1842 branches, and the number of Mobile and Internet Banking users has exceeded 6.6 million and 37k, respectively. The number of active ATM card users and CBE birr users reached more than 8.3 million and 17 million, respectively [4]. Electronic payments have revolutionized business processing by reducing paperwork, transaction costs, and labor costs. As a result of CBE's aggressive expansion, the volume of transactions and the number of customers is increasing rapidly. This has generated a massive amount of data; which CBE manages using an Enterprise Data Warehouse (EDW). An EDW is a database, or group of databases, that collects and centralizes corporate data from many sources and applications and makes it accessible for analytics and use throughout the enterprise.

1.2 Statement of the Problem and Research Question

According to [6], specific use cases and areas of business application should serve as the foundation for considerations and performance standards for Customer Lifetime Value models in the banking context.

Research has been conducted regarding data mining applicability on customer segmentation, customer relationship management, and customer churn prediction for banks in Ethiopia. Yinebeb [7] worked on “cluster analysis for customer segmentation at the Commercial Bank of Ethiopia.” Birhane [8] on the other hand worked on “Bank customer churn prediction: the case of

Commercial Bank of Ethiopia." Wakgari [9] also conducted research on the "Application of data mining techniques for effective customer relationship management of microfinance in the case of wisdom microfinance." Their results show that the problem of customer relationship management can be solved using data mining technology. While not explicitly developing a CLV prediction model, the aforementioned study provides the data foundation and useful insights regarding customer behaviour and segmentation that can increase model accuracy and usefulness. However, to the best of our knowledge, there is a lack of local studies on the application of data mining techniques for customer lifetime value prediction in the banking industry in general and CBE in particular. The Ethiopian banking sector has a limited understanding of customer lifetime value. Existing research focuses on churn prediction and customer segmentation, leaving out the critical feature of CLV analysis. This creates a gap or lack of knowledge in understanding the key factors contributing to predict customer lifetime value for CBE. Identifying these determinant attributes can help the company develop targeted marketing strategies and improve customer acquisition and retention.

Currently, CBE is using statistical analysis as a strategic tool to identify potential or high-value customers among its customer base. This approach involves extracting relevant data from the data warehouse and using it in conjunction with the CRM system to precisely identify customers with large deposit amounts, indicating potential profitability for the bank. While this strategy has been efficient in identifying valuable customers, current research [10] indicates that factors other than customer deposits could play an important role in determining a customer's value. These additional criteria include demographic information, spending habits, and transactional details. By merging these varied elements into its analytical models, CBE can acquire a better understanding of each customer's behavior and preferences. Such detailed analysis can help CBE design more targeted marketing strategies that better meet the demands and interests of individual customers. In summary, by expanding its focus beyond deposit levels when measuring customer value, CBE can improve its ability to target specific segments more efficiently.

The proposed data mining approach aims to better understand CBE customers by analyzing their demographics and transactional data. By incorporating various attributes, the model can predict the lifetime value of each customer, providing a more detailed view. This approach recognizes that valuable customers are determined not only by their deposits but also by other factors. To effectively estimate CLV, determine the key factors driving customer value in the Ethiopian

banking sector. Analyzing customer data and incorporating insights from previous studies helps to discover the most important elements for predicting CLV. Understanding these essential factors is critical to creating an effective CLV prediction model specific to the CBE's customer base. In this study, we propose using a supervised machine learning approach with attributes such as customer demographics and transactional data to predict the lifetime value of CBE customers. The main goal of this research is to apply data mining techniques to predict the CLV of CBE customers, leading to more effective CRM.

Customer lifetime value (CLV) prediction is crucial for businesses to identify and retain their most valuable customers. There are many machine learning algorithms available for CLV prediction, but it is unclear which one is more suitable for CBE. This is due to the fact that no research has precisely examined the effectiveness of various machine learning algorithms for CLV prediction at CBE. Most existing studies have used data from other countries and industries. This model fills a gap in local research by offering a data-driven way to predict customer lifetime value. This study seeks to assess and compare the performance of appropriate algorithms for CLV prediction, taking into account the features of the CBE's customer data. Finding the most effective algorithm ensures that the CLV prediction model produces accurate and dependable data, allowing the CBE to make educated decisions based on customer value. The study presents a data-driven approach to CLV prediction by defining key variables, selecting an effective algorithm, and evaluating its performance. This would help CBE to identify the most suitable algorithm for their needs. Therefore, the study will try to answer the following research questions:

- ✓ What are the significant attributes that help to predict the lifetime value of CBE customers?
- ✓ Which algorithm best suits to customer lifetime value prediction modeling at CBE?
- ✓ To what extent the selected algorithm works in the CLV prediction?

1.3 Objective

1.3.1 General Objective

The general objective of this study is to investigate the application of data mining techniques for predicting customer lifetime value to improve customer relationship management at the Commercial Bank of Ethiopia.

1.3.2 Specific Objective

To achieve the general objective, the research has the following specific objectives:

- ✓ To identify and analyze relevant customer data variables that impact customer lifetime value.
- ✓ To prepare and preprocess the data sets' attributes for the prediction model.
- ✓ To identify and find the attributes that can help to predict customer lifetime value.
- ✓ To select appropriate data mining tools and algorithms to be used for developing the CLV prediction model.
- ✓ To construct a customer lifetime value prediction model using the selected data mining tools and algorithms.
- ✓ To evaluate the proposed customer lifetime value prediction model using performance measures.
- ✓ To report the finding of the study for the upcoming research area in the field

1.4 Significance of the Study

The main advantage of this study is that it contributes to both model building and practice improvement in customer lifetime value prediction. In this study, CBE is the main beneficiary. Mainly, the study is significant in managing customers by knowing how much is worth and gaining business advantage from the customer. Likewise, in making more informed decisions about how to allocate their resources. The banks can target their marketing and retention efforts on those customers who are likely to be high-value, long-term customers while simultaneously taking measures to reduce attrition among lower-value customers. Moreover, the findings of this study allow the business to understand their customers and identify those who are most valuable. The information can then be used to create a more targeted marketing campaign and provide a better customer experience, which will lead to customers staying with the business for longer. In

addition, this study can also contribute to the study area of building customer lifetime value prediction models of a data mining application for banking industries to maximize the value of customers in their marketing strategies.

However, by understanding the factors that contribute to CLV, the bank can better understand what drives their customers to do business with them. This information can be used to develop more effective marketing and product development strategies. The bank can use CLV prediction to segment their customers and develop targeted marketing campaigns, which helps increase revenue from existing customers.

1.5 Scope of the Study

The proposed research aims to develop a model for predicting the lifetime value of CBE customers based on the collected dataset. We will use data-mining technologies to predict only the lifetime value of CBE customers. We have collected data and used data-mining techniques to build models for predicting customer lifetime value. The process includes steps such as understanding the business, understanding the data, preparing the data, building models, and evaluating them with test datasets. For this study, we analyzed historical transactional data from one year. We chose this timeframe because extracting more than one year of transactional data requires additional computational power. The bank has processed billions of transactions, resulting in a large table that poses challenges when extracting information from it. Longer timeframes may give more accurate predictions but require more computational resources. Additionally, deploying the models was not part of our scope since our focus is on developing and evaluating them. Moreover, deployment requires coordination of resources, a collaborative effort from people, organizational procedure, and technology.

1.6 Organization of the Thesis

There are five chapters in this research report. The first chapter covers the basic overview of the study, including the study's background, the Commercial Bank of Ethiopia's background, the problem statement, the general and specific research objectives, the study's scope, and the study's importance. A conceptual assessment of the literature on data mining technology and related works on data mining applications for predicting customer lifetime value is covered in the second chapter. The study's methodology is covered in chapter three. The focus of chapter four is

experimentation and analysis, with related discussions and findings. In Chapter five, which gives the study's conclusion and offers suggestions for additional research, the topic of conclusion and recommendations is finally covered.

CHAPTER TWO

LITERATURE REVIEW

2.1 Overview

A thorough literature assessment is essential for knowing the present level of knowledge about customer lifetime value (CLV) prediction, particularly in the Ethiopian banking sector. This review will cover the following major aspects: CLV Prediction in the Banking Context review existing studies on CLV prediction models used by banks globally. Data Mining Applications in Ethiopian Banking: Identify research that use data mining for customer analysis in Ethiopian banks. Research Gap and Opportunity is then discussed by analyzing research limitations in Ethiopia, including a lack of studies on CLV prediction.

A commercial bank is a financial institution that serves customers with basic banking services. Accepting deposits, making loans, and providing investment products are examples of these services. A bank that accepts deposits, makes business loans, and provides necessary investment products for profit. It is typically involved with controlling withdrawals and accepting deposits, as well as offering short-term loans to individuals and small enterprises. Customers generally use these banks to check, save, deposit, and secure loans against their homes or other assets.

2.2 The Banking Sector

The number of banks operating in Ethiopia reached 30 by the end of June 2022. These banks opened 1600 new branches in 2021/22, increasing the total number of branches to 8,944 from 7,344 a year ago. Consequently, the population-to-bank-branch ratio increased to 12,000 people per branch. Addis Ababa accounts for approximately 32.7 percent of all bank branches. Due to the opening of 1,600 new branches, private banks' share of total deposit mobilization increased to 47.7 percent from 45.7 percent in 2022. Due to its extensive branch network, CBE alone mobilized 52.3 percent of total deposits [12].

Over the last decade, Ethiopian banks have played a significant role in widening financial access, enhancing national savings, and funding large public and private enterprises. Banks have recently emerged as a major source of employment, income, and taxes. Despite these contributions, progress in some critical areas remains restricted, especially when compared to

other countries and in light of local private sector needs. Banking in Ethiopia is a diverse industry that is dominated by a government-owned corporation that controls roughly two-thirds of the market [13].

2.3 Commercial Bank of Ethiopia

As mentioned in Chapter One, The Commercial Bank of Ethiopia was established in 1942 as the State Bank of Ethiopia. The CBE was formally constituted as a share company in 1963. The CBE and the privately owned Addis Ababa Bank merged in 1974. Since then, it has made a substantial contribution to the nation's growth. CBE now has over 39.9 million account holders in its 1842 branches, while the number of Mobile and Internet Banking customers has surpassed 6.6 million and 37k, respectively. There are almost 8.3 million active ATM card holders and 17 million CBE Birr users. CBE is known as one of the predominant banks operating in Ethiopia. CBE is credited with being a catalyst for Ethiopia's economic development and for being a pioneer in bringing modern banking to the nation. Additionally, it was the first bank in Ethiopia to offer residents access to ATMs, to offer Western Union money transfers, and to partner with more than 20 money transfer companies at the moment [4].

CBE offers deposit services such as savings accounts, current-checking accounts, and diaspora accounts, as well as Credit Services, International Services such as trade services, forex services, money transfers, and respondent banking, and Payment Services such as Internet Banking, Card Banking, Mobile Banking, and Card Local Transfer. In the past 80 years, CBE has supported the development projects of the country by supporting the overall economic activity through financing, and access to banking services and is a leader in expanding and promoting technology-based banking services. It has been playing a role in the rapid development of the country in the public and private sectors. As the number and scale of development projects are growing significantly, the bank is responding accordingly. At the end of the second quarter of 2022/23, banks' total assets reached birr 1.2 trillion and their capital increased to 60 billion [14].

Statistical analysis is currently used by CBE to identify potential or high-value customers. Based on customer deposits, the CRM identifies high-value customers. This strategy assists the bank in identifying customers who are more likely to be beneficial. Customers are divided into three groups based on the value and volume of their transactions: retail customers with a yearly average transaction of less than Birr 100,000.00, business customers with a yearly average

transaction of between Birr 100,000.00 and Birr 1,000,000.00, and premium customers with a yearly average transaction of Birr 1,000,000.00 or more. There are no other ways to estimate the lifetime value of customers in CBE. This study serves as an input for further research in this area.

2.4 Customer Relationship Management

As cited in [15], Four simple framework aspects define Customer Relationship Management (CRM): Know, Target, Sell, and Service. CRM necessitates familiarity with and understanding of the company's markets and customers. This includes acquiring thorough customer intelligence in order to identify the most profitable customers and removing those who are no longer worth targeting. CRM is one of the most effective approaches and tools for increasing the customer base and thus surviving in this competitive environment. CRM with customers is now used by the banking sector to obtain customer databases, customer satisfaction levels, customer loyalty, long-term service, and customer retention, as well as to identify profitable customers for their banks [15].

Customer relationship management is primarily concerned with understanding customers and profitability as well as retaining profitable customers. The CRM model includes systems that support and manage four types of activities. The first stage involves attracting new customers and retaining profitable ones. Customer preferences are identified in the second stage. Customers are differentiated based on their preferences, and strategies for interacting with them are developed, taking into account the differences between various customer segments. In the next stage, the organization offers customized solutions to customers based on their needs and preferences to motivate them to take action. Finally, the organization completes transactions and provides customer service. Although these are different stages of the CRM process, they are not exclusive. The various stages of the process frequently overlap, and the process itself is iterative [16].

This study addresses the first stage of CRM activities to retain profitable customers. This study's goal is to explore the applicability of data mining techniques in the early stages of CRM activities that companies engage in to keep hold of their most valuable customers. In today's business environment, customer retention is more important than ever. To build strong customer relationships that lead to increased loyalty and repeat business, it is critical to understand the first stage of CRM operations.

2.5 Customer Lifetime Value

The relationship between a customer and an organization that allows the organization to benefit more from its customers is referred to as the customer's lifetime value. The present value of all future revenues earned from a customer throughout his or her engagement with the organization is called CLV. This is comparable to the discounted cash flow method in finance [17].

CLV is a metric that measures how valuable a customer is to a business over the long term. It is an important metric for businesses to track because it can help them to identify and retain their most valuable customers, and to optimize their marketing and sales efforts. CLV, which assesses a customer's profit-generating potential or value, is increasingly being viewed as a touchstone for operating the CRM process in order to give attractive services to and retain high-value customers while optimizing business profitability. Customer lifetime value modeling covers a wide range of applications, including customer-specific services and offers, detecting and managing unprofitable customers, customer segmentation, marketing, pricing, and promotional analysis [18]. As cited in [19] CLV has received a lot of attention in recent years in CRM research. It calculates a customer's worth to a company by taking into account expected future transactions and expenses. The CLV formula comes in a variety of modifications. The simplest formula for CLV calculation is shown below.

CLV = Average Purchase Value * Purchase Frequency * Average Customer Life Span [20].

Based on the above form, we have calculated the estimated CLV using the formula:

CLV = Average Transaction Amount * Transaction Frequency *Average Customer Life Span.

The above formula was used to compute CLV for all customers. This will be utilized as the Prediction Target variable. After calculating each customer's CLV, we used data mining to identify additional factors influencing customer lifetime value beyond the calculated attributes. Data mining enables predictive modeling, which creates models that predict individual customer future CLV based on historical data and criteria other than average transaction amount, frequency, and lifespan.

2.6 Data Mining

Finding patterns and information in vast amounts of data is a technique known as data mining. Data sources include, for instance, databases, data warehouses, the web, other information repositories, and data that is dynamically provided into the system [21]. "Data mining" refers to the technique of obtaining meaningful information from vast volumes of data. Many additional terminology have been used to interpret data mining, including knowledge mining from databases, knowledge extraction, data analysis, and data archaeology [22]. Data mining is useful for identifying patterns, forecasting, and discovering knowledge across a variety of business disciplines. Data mining techniques and algorithms, such as classification and clustering, aid in the discovery of patterns in order to determine future business trends. Data mining has a wide range of applications in practically any industry that generates data, which is why it is regarded as one of the most important frontiers in database and information systems research, as well as one of the most promising multidisciplinary advances in Information Technology [3].

2.7 Data Mining Process Methodologies

The Knowledge Discovery Database (KDD) process model, the Industry Standard for Data Mining (CRISP-DM), and SEMMA are the three most popular data mining process models. Data mining professionals and academics primarily employ these three models. The three data-mining approaches are discussed in the next section.

2.7.1 Knowledge Discovery in Database (KDD)

The KDD process use DM approaches to extract what is deemed knowledge based on the specification of measurements and thresholds, using a database and any necessary preprocessing, subsampling, and database transformation. The first stage of KDD is selection. This stage entails constructing a target dataset or focusing on a collection of variables or data samples on which discovery will be performed. The second stage is preprocessing. This step included target data cleaning and preparation in order to obtain consistent data. Transformation is the third stage. This level entails data manipulation using dimensionality reduction or transformation methods. Data mining is the fourth stage. This stage involves seeking for patterns of interest in a certain representational form, depending on the data-mining goal (typically prediction). Interpretation/Evaluation is the fifth stage. This involves interpreting and evaluating the mined patterns [23]. figure 2.1 shows the steps in the KDD process.

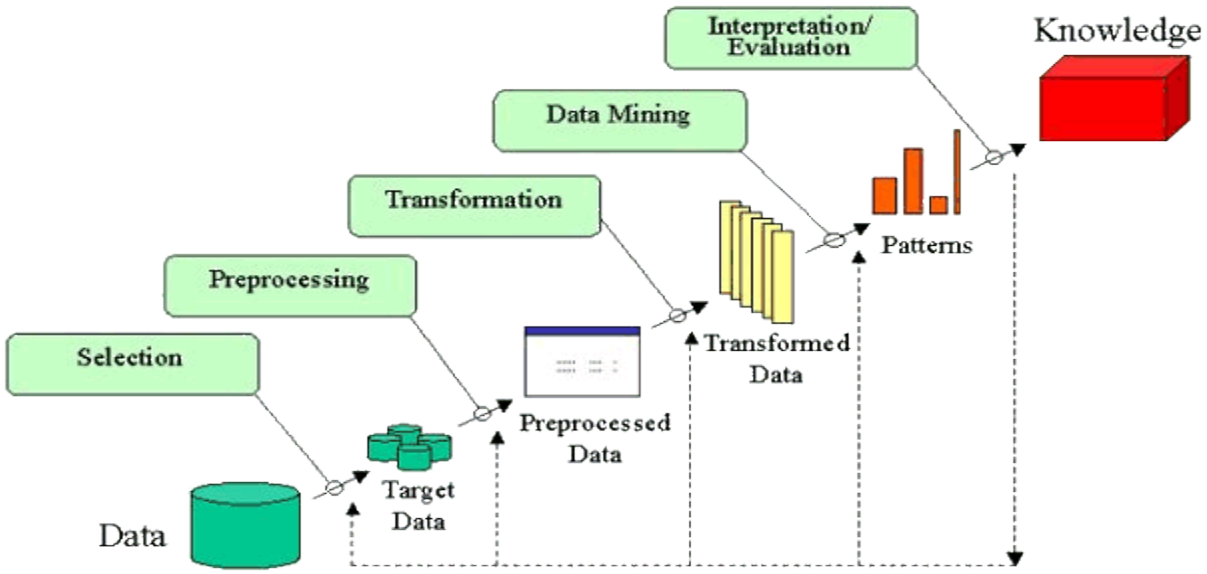


Figure 2.1 Steps of the Knowledge Discovery in Database process adopted from [24].

2.7.2 CRISP-DM

CRISP-DM, or Cross-Industry Standard Process for Data Mining, is a six-step process for extracting knowledge from data. It is a flexible framework that can be adapted to different data mining projects and industries. The first step of CRISP-DM is business understanding. This involves understanding the business goals and objectives of the project, as well as the data that is available. The goal of this step is to define the data mining problem and develop a plan to achieve the desired results. The second step is data understanding. This involves exploring the data to get a better understanding of its characteristics, including its quality, format, and content. The goal of this step is to identify any potential problems with the data and to develop a plan to address them. The third step is data preparation. This involves cleaning and transforming the data into a format that is suitable for data mining. This may involve removing outliers, correcting errors, and filling in missing values. The goal of this step is to create a high-quality dataset that is ready for analysis. The fourth step is modeling. This involves applying data mining techniques to the data to discover patterns and relationships. There are many different data mining techniques available, and the best approach to use will depend on the specific problem that is being solved. The goal of this step is to develop a model that can be used to make predictions or decisions about the data. The fifth step is evaluation. This involves assessing the performance of the model on a held-out test set. This helps to ensure that the model is generalizable and that it will not over fit the training data. The goal of this step is to identify the best model for the problem and to

make any necessary adjustments. The final step of CRISP-DM is deployment. This involves making the model available to users so that it can be used to make predictions or decisions about new data. This may involve integrating the model into a software application or creating a web service that exposes the model to users. The goal of this step is to ensure that the model is used in a way that delivers value to the business [21].

CRISP-DM is a valuable tool for data scientists and other professionals who work with data. It provides a structured approach to data mining projects and helps to ensure that projects are completed successfully [23]. CRISP-DM is the most widely used framework for data science projects. This was first published in 1999. It provides a natural description of a data science lifecycle (workflow in data-focused projects) [25]. Figure 2.2 shows the steps of the CRISP-DM process model.

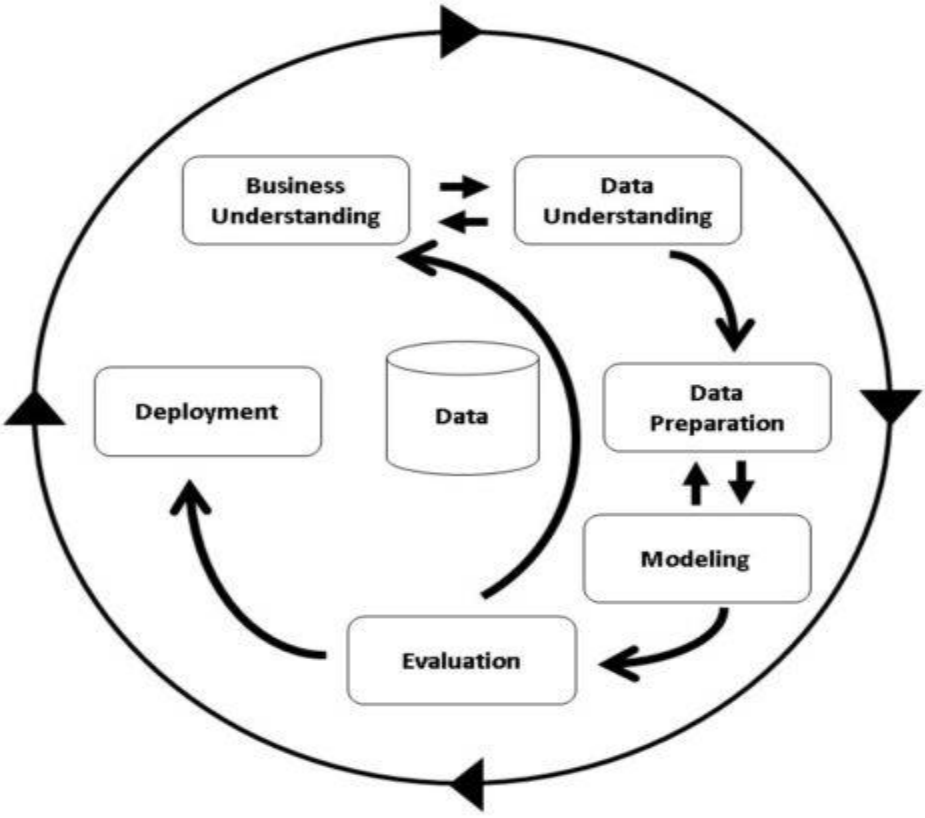


Figure 2.2 Steps of the CRISP-DM process model adopted from [5].

2.7.3 The SEMMA Process Model

The SEMMA process is a five-step data mining process developed by the SAS Institute. It stands for Sample, Explore, Modify, Model, and Assess. Stage one is Sample: This step involves selecting a representative sample of the data to be used in the data mining project. If the dataset is too large to process entirely, this step becomes necessary. Next is Explore: This step involves exploring the data to understand its characteristics and identify any potential problems. Data visualization techniques can be used to look for patterns and trends in the data. The third stage is Modify: Here, we prepare the data for modeling. This includes cleaning it up, removing outliers, and transforming it into a suitable format for our chosen modeling technique. Then comes Model: In this step, we apply various data mining techniques (such as classification, regression, and clustering) to build a model that can predict our desired outcome. Lastly is Assess: We evaluate the performance of our model on a held-out test set. This ensures that our model isn't overfitting or biased towards just training data; it needs to be generalizable [21].

Although the SEMMA technique is independent of the user's preferred DM tool, it is connected to the SAS Enterprise Miner software and makes the claim to help the user implement DM applications. SEMMA offers a straightforward procedure for creating and maintaining DM projects in an appropriate and logical manner. As a result, it offers a framework for his formation, evolution, and invention, helping to propose business solutions and establish DM business objectives [23]. figure 2.3 shows the steps of the SEMMA process model.

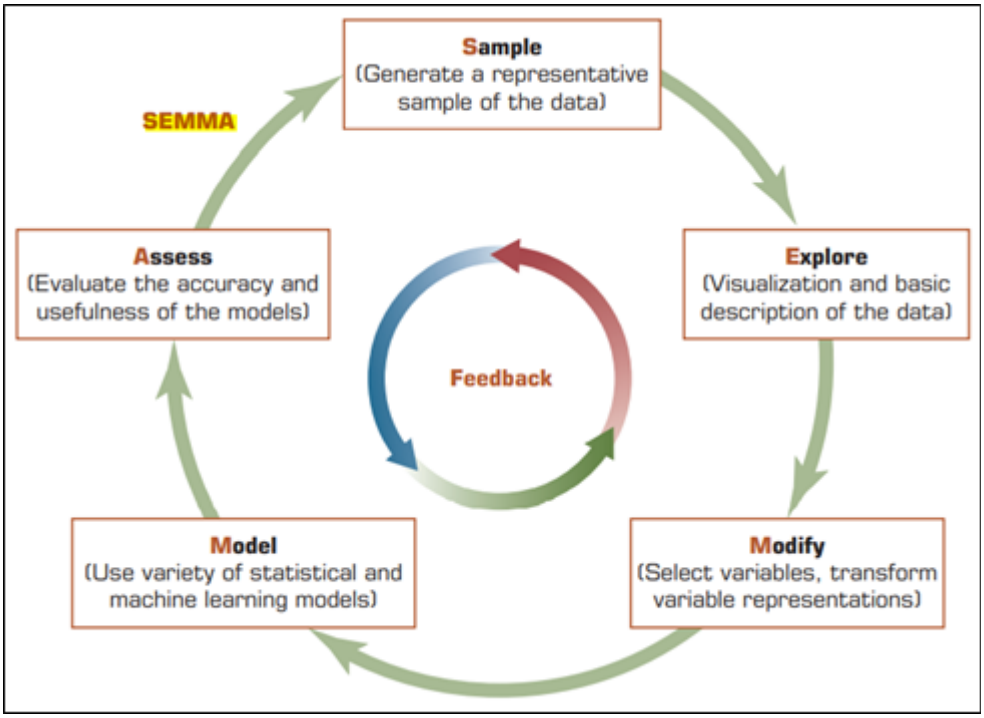


Figure 2.3 Steps of the SEMMA Process Model adopted from [5]

Table 2.1 Data Mining Process Methodologies Comparison

DM Process Methods	KDD	SEMMA	CRISP-DM
Process steps	Pre KDD	----------	Business understanding
	Selection	Sample	Data understanding
	Preprocessing	Explore	Data preparation
	Transformation	Modify	
	Data mining	Model	Modeling
	Interpretation Evaluation	Assessment	Evaluation
	Post KDD	----------	Deployment
Usage	Rarely used in industry.	Commonly used in the pharmaceutical industry	Used in a variety of industries
Acceptance	Less widely known	Less widely accepted	Widely accepted

CRISP-DM has proven to be the most widely used and accepted of the three process models discussed above [26],[27]. CRISP-DM and SEMMA are two popular data mining frameworks

that are both based on the KDD process. CRISP-DM is a cross-industry standard that can be used for data science projects in any domain. It also provides a reliable framework for project scheduling and management.

2.8 Data Mining Tasks

Data mining tasks can be broadly classified into four categories. Classification is putting data into predefined groups. Clustering is grouping similar data together. Regression is finding a function that models the relationship between two or more variables. Association rule learning is discovering relationships between different variables. These four classes of data mining tasks can be used to solve a wide variety of problems in different industries [22].

2.8.1 Classification

Classification is a data mining technique that uses a set of pre-classified examples to develop a model that can classify new data points. It is the most commonly used data mining technique, and it has a wide range of applications, including fraud detection, credit risk assessment, and customer segmentation. Classification algorithms typically work by first learning from a set of training data, which consists of data points that have already been classified into different categories. The algorithm then uses this information to build a model that can be used to classify new data points. Classification is a powerful data mining technique that can be used to solve a wide variety of problems. It is relatively easy to understand and implement, and it can be used with a variety of different data types. There are many different classification algorithms available, but some of the most common ones include: decision trees, neural networks, Bayesian classifiers and support vector machines (SVMs) [28].

2.8.2 Clustering

A data mining technique called clustering brings together comparable data points. It can be used to locate thick and sparse areas in the data space as well as to find patterns and correlations in the data. As a preprocessing step for classification tasks like predicting gene function or customer segmentation, clustering can also be employed. The most popular clustering techniques are Partitioning methods, Hierarchical agglomerative (divisive) methods, Density-based methods, Grid-based methods, and Model-based approaches [28].

2.8.3 Regression

A data mining method that can be used to forecast results is regression. Modeling the correlation between a set of independent variables (predictor variables) and a dependent variable (target variable) is how it operates. In data mining, the dependent variable is what we wish to predict, while the independent variables are often well-known qualities. Numerous issues can be resolved using regression, including forecasting sales volumes, stock prices, and product failure rates. Many situations in the real world, however, are not accessible to straightforward forecasts. These issues frequently involve numerous predictor factors and are complicated. More sophisticated regression approaches, including logistic regression, decision trees, or neural networks, may be needed to forecast future values for these issues. Similarly, these model types are commonly employed in categorization problems [28].

To categorize an issue as supervised or unsupervised learning, machine learning algorithms can be utilized. In supervised learning, the algorithm is trained on a set of labeled data in which both the input characteristics and the output target variable are known. The algorithm then develops the ability to anticipate the target variable for the output given fresh input information. When learning without supervision, the algorithm is trained on a collection of unlabeled data, where the input characteristics are known but the output target variable is unknown. A pattern- and relationship-finding algorithm is then trained to find patterns in the data.

This study uses supervised learning algorithms to determine which ones are most advantageous for examining customer behavior and estimating lifetime value. The total revenue that a customer is anticipated to bring in throughout the course of their association with the business is measured by customer lifetime value. This research will examine the machine learning techniques of linear regression, decision tree regression, and random forest regression. These algorithms were chosen because prior studies have demonstrated their efficacy in predicting client lifetime value.

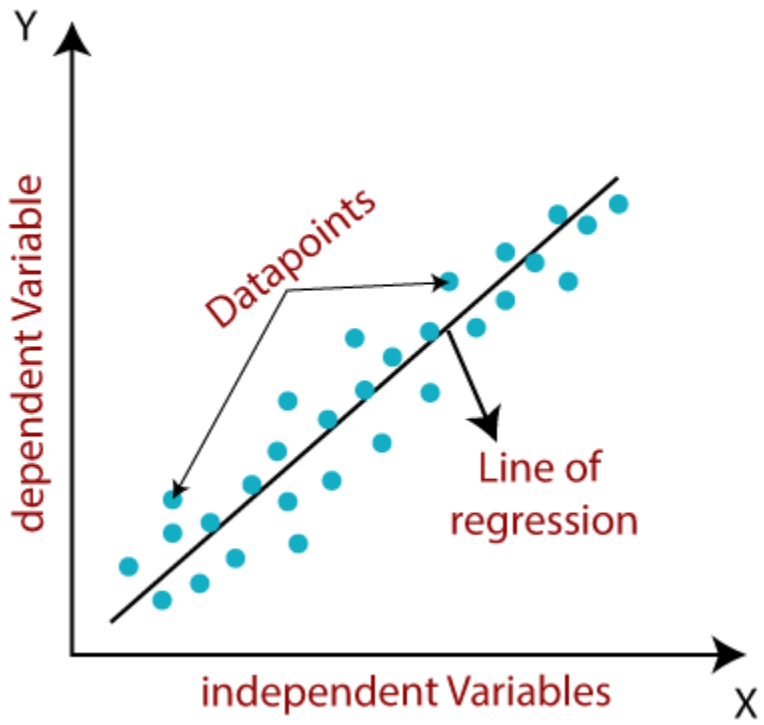
2.8.3.1 Linear Regression

Linear regression is a statistical method that is used to model the relationship between a dependent variable (also known as the target variable or response variable) and one or more independent variables (also known as predictor variables or input variables). It is one of the most well-known and widely used statistical methods. Regression was first developed by Sir Francis

Galton and Karl Pearson in the 1920s and 1930s to study the inheritance of traits in sweet peas. Since then, it has become the go-to statistical method for characterizing the relationships between explanatory (input) and response (output) variables in a wide range of fields, including science, engineering, business, and economics. Regression is a relatively simple statistical technique, but it is very powerful. It can be used to:

- ✓ Identify and measure the strength of relationships between variables.
- ✓ Make predictions about the future.
- ✓ Test hypotheses about the relationships between variables.

Regression is predicated on the assumption that a linear function may adequately represent the relationship between the dependent variable and the independent variables. However, connections between variables in the real world frequently deviate from linearity. More intricate regression models can be applied in these circumstances. A variety of issues can be resolved using the flexible tool of linear regression. It is an effective instrument for understanding and predicting the external environment [5],[29].



Linear regression can be represented mathematically as $y= a_0+a_1x+\varepsilon$

Where Y stands for Dependent Variable (Target Variable)

X stands for Independent Variable (predictor Variable)

a0 = the line's intercept (Gives an additional degree of freedom)

a1= Linear regression coefficient a1 (scale factor to each input value).

ε = random error

2.8.3.2 Decision Tree Regression

It has three different sorts of nodes and is organized like a tree. The first node that represents the complete sample and can be further divided into nodes is known as the Root Node. Interior nodes stand in for a data set's characteristics, while branches stand in for its decision-making procedures. The result is finally represented as Leaf Nodes. When dealing with issues that call for decisions, this algorithm is really helpful. By responding to True/False questions, a given data point is moved entirely through the tree until it reaches the leaf node. The average value of the dependent variable in that specific leaf node serves as the foundation for the final projection. The Tree can predict an acceptable value for the data point after several iterations [30].

A machine learning model called a decision tree learns to generate predictions by adhering to a set of decision rules. It is a supervised learning model, meaning that the input features and the goal variable for the output are both known before the model is trained on a collection of labeled data. For new input features, the decision tree learns to forecast the output target variable. Each node in a decision tree's representation is a choice, and the structure resembles a tree. The root node, which is the tree's top node, is a representation of the complete dataset. The leaf nodes reflect the anticipated values for the target variable, whereas the inside nodes represent other dataset characteristics. A decision tree makes decisions at each node based on the values of the input features as it moves from the root node to the leaf node to produce a forecast. The value connected to the leaf node that the tree reaches is the expected value. Classification, regression, and anomaly detection problems can all be resolved using decision trees, a potent machine learning technique. They are a common choice for many machine learning applications since they are also quite simple to understand and interpret [31].

2.8.3.3 Random Forest Regression

Random forest regression is a Tree-based technique for decision-making that makes use of many Decision Trees' properties. As a result, it can be referred to as a "forest" of trees, hence the term "Random Forest." This algorithm is called "Random" because it is a forest of "Randomly created Decision Trees." [30].

The random forest, as its name suggests, is made up of a sizable number of distinct decision trees working together as an ensemble. The random forest generates class predictions for each tree, and our model predicts the class that receives the most votes [32].

An approach for supervised machine learning called random forest regression is built on the idea of ensemble learning. A method called ensemble learning mixes different machine learning models to enhance the performance of the entire model. It functions by constructing a number of decision trees, which are then averaged to produce a single prediction. The overfitting that might happen with individual decision trees is lessened by this method. The steps that are commonly taken in order to construct a random forest regression model are as follows:

Step 1: Choose a random sample of the training data.

Step 2: Build a decision tree on the sampled data.

Step 3: Repeat steps 1 and 2 for a specified number of times.

Step 4: To make a prediction for a new data point, average the predictions from all of the decision trees in the forest.

Random forest regression is a powerful and versatile algorithm that can be used to solve a wide variety of regression problems. It is particularly well-suited for problems with complex relationships between the input and output variables [33].

A random forest predicts via averaging after fitting numerous classification decision trees to subsamples of the dataset. Because several instances of the data were used, predictive accuracy is raised and overfitting is avoided. The model's accuracy can be improved by changing the parameters. Random Forests are appealing from a computational standpoint because they

naturally handle both regression and (multi-class) classification; they are relatively fast to train and predict; they rely on only one or two tuning parameters; they have a built-in estimate of generalization error; and they can be used directly for high-dimensional problems [34].

2.8.4 Association Rule

An approach to data mining called association rule mining looks for recurring patterns among items in huge databases. Business decisions like catalog design, cross-marketing, and customer shopping behavior research can be aided by this kind of finding. Algorithms for mining association rules must be able to generate rules with confidence levels below one. This means that even while the rules aren't perfect, they are still likely to be accurate a sizable portion of the time. For a given dataset, there are typically many different association rules that may be used, and the majority of these rules are frequently ineffective or useless. Therefore, it is important to develop efficient algorithms that can identify the most useful rules. There are many possible association rules for a given dataset, so it is important to use efficient algorithms to identify the most useful rules. Association rules can be used to improve product recommendations, design more effective marketing campaigns, and better understand customer behavior [28].

2.9 Model Performance Evaluation

This section describes the evaluation approach used to assess the performance of the customer lifetime value prediction models. Evaluation metrics are quantitative measures that are used to analyze the accuracy and precision of a model's predictions. Evaluation metrics reveal how well a model performs on a particular dataset. The performance of regression models is assessed and reported using R^2 and RMSE.

2.9.1 R Squared (R^2)

A well-fitting regression model is one that produces predictions that are close to the observed values. There are a number of statistical indicators that can be used to assess the fit of a regression model, all of which are based on the sum of squared errors (SSE), which measures how far the data points deviate from the mean and the model's predictions. Different combinations of these two values can provide insights into how well the regression model performs compared to a simple mean model.

One of the most widely used and interpretable measures of regression model fit is the Coefficient of Determination or R-squared. R^2 ranges from 0 to 1, with higher values indicating a better fit. An R^2 of 0 indicates that the model does no better than predicting the mean, while an R^2 of 1 indicates that the model perfectly fits the data. In practice, R^2 values are typically between 0 and 1, with higher values indicating better predictive performance. Overall, a well-fitting regression model is one that produces predictions that are close to the observed values and has a high R-squared value [5].

2.9.2 Root Mean Squared Error

Root mean squared error (RMSE) is a statistical measure of how close a regression model's predictions are to the actual values. It is calculated by taking the square root of the mean squared error (MSE). This means that the RMSE is in the same units as the predicted values, which makes it easier to interpret.

RMSE is a commonly used metric for evaluating the performance of regression models because it is easy to calculate and interpret. It is also a good measure of overall model fit, as it penalizes large errors more than small errors.

The formula for calculating RMSE is stated as follows:

$$\text{RMSE} = \sqrt{\left(\frac{1}{n} * \sum_{i=1}^{n} (\text{yi} - \hat{y}i)^2\right)} \quad (1)$$

Where n is the number of data samples, yi is the dataset's i^{th} expected value, and $\hat{y}i$ is its i^{th} forecast value. A lower RMSE indicates a better model fit, as it means that the model's predictions are closer to the actual values [35].

2.10 Related Works

Research on data mining techniques' applicability in the financial industry is getting more popular. Among them, research conducted locally or globally that is related to customer lifetime value is discussed below.

2.10.1 Global Studies

Several studies have been conducted on customer lifetime value modeling. Some of these studies have been reviewed and discussed below.

The research entitled “A two-stage machine learning approach for modeling customer lifetime value in the Chinese airline industry” worked on customer lifetime value modeling. In the context of the study, the two-stage strategy refers to a machine learning approach that involves two stages. Three classic machine learning approaches (logistic regression, gradient boosting decision tree, and neural network models) are utilized to model CLV as high value passenger or low-value passenger in the first step. The results from the first stage are fused in the second stage using an innovative form of evidence reasoning (ER) modeling to improve prediction accuracy. When faced with uncertainty or conflicts in several pieces of information, the ER technique offers the advantage of fusing data and providing categorization conclusions. Overall, the two-stage strategy improves CLV assessment accuracy by overcoming the limitation of relying on a single model. The primary goal of this study was to overcome the limitations of the existing CLV model. To improve prediction accuracy, this study integrates internal and external data into CLV models and introduces the evidence reasoning (ER)-based model fusion approach. They used data from Airline X's 107,538 passengers. The dataset includes 327 attributes about four aspects of the travel experience: passenger profile, travel history, route characteristics, and loyalty program status. They use a case study of a Chinese airline to demonstrate a two-stage approach. They conducted a discrete analysis using logistic regression, gradient boosting decision tree (GBDT), and neural networks to estimate the CLV for airlines. The accuracy of the logistic regression model, GBDT, neural network model, and ER-based model fusion was 0.775,0.818, 0.810, and 0.834 respectively. For the analysis of the data, Python 3.0 was utilized. The proposed two-stage approach can assist airlines in identifying high-value passengers. As a result, airlines can create individualized CRM strategies to strengthen and retain customer connections [1].

The other research conducted was entitled “Customer Lifetime Value Analysis Based on Machine Learning.” This article reviews machine-learning models for CLV analysis and suggests future research directions. In this study data from 8099 samples with 23 features were collected from Kaggle and analyzed using four different machine learning methods such as linear regression, support vector machine, random forest, and neural network. Among the four machine learning models presented in the article, Random Forest is the best for assessing CLV since it has the lowest MSE and MAE values and the highest R^2, 0.70. According to the correlations between

features, CLV is generally affected by monthly premium auto, total claim amount, and coverage. Machine learning models have high precision and random forests perform best [36].

The thesis entitled “Customer Lifetime Value Prediction and Segmentation using Machine Learning” was another. The goal of this project is to analyze a company's customer sales data and predict the customer’s lifetime value. Customer segmentation was performed to determine focus groups based on the predicted CLTV. The goal of this project is to provide a guide for marketing decision-making as well as future marketing strategies and plans by utilizing machine learning models to predict customer lifetime values and segmentation. This project was conducted using the CRISP-DM methodology. The dataset includes all sales transactions from January 1, 2020, to December 31, 2020. with Attribute Details: (Customer ID, Invoice Number, Item Number, Quantity, Invoice Date, Unit Price, Customer Category). Implementing CLTV-based strategies will provide valuable insights into how to improve the customer experience. Based on the prediction accuracy, the Gradient Boosting Regression model was considered for CLTV prediction in this project with an R^2 value equal to 0.84. The CLTV prediction calculates the total customer value for each customer. This will be beneficial to the marketing team because it provides them with specific customers to target. However, to strategize and build a marketing plan, it is critical to create clusters or segments to make it more actionable. K-means clustering was used to form groups based on the predicted CLTV [20].

The thesis entitled “Modelling Customer Lifetime Value in the Retail Banking Industry” is also the other. The purpose of this thesis was to shed light on the proper modeling of Customer Lifetime Value in the retail banking sector and to better understand the factors that must be taken into account throughout the modeling process. First, performance standards for models of Customer Lifetime Value in the retail banking sector were established through literary research and interviews with SEB. Six general modeling approaches RFM, Probability, Econometrics, Persistence, Diffusion/Growth, and Computer Science were then subjected to these requirements. Based on the examination, it was determined that the econometric and computer science approaches were suitable for further study. This was done by putting two models into practice and evaluating how well they performed as examples of each strategy. Particularly, an econometric model based on Markov chains and a computer science model based on the random forest technique were chosen. According to the findings, both approaches would be suitable for the retail banking sector, however an econometric approach might have a larger interpretability

advantage and a computer science approach might have a better predictive accuracy advantage. The results show that specific needs for Customer Lifetime Value models in the context of retail banking should be based on a particular use case and area of business application [6].

2.10.2 Local Studies

"Cluster analysis for customer segmentation in commercial bank of Ethiopia." This research examines the application of data mining techniques to customer segmentation in the commercial bank of Ethiopia using the cross-industry standard process for data mining (CRISP-DM) methodology. The data set includes 216,721 customer accounts and 20 attributes that include customer demographic, account, usage, and transaction information. The researcher uses Rapid Miner to accomplish the desired data preprocessing and mining task. This research compares three clustering algorithms: Agglomerative, DBSCAN, and K-means. As a result, K-Means outperformed the clusters in terms of execution time and average Sum of Squared error. Furthermore, K-means is one of the most widely used and simple clustering algorithms, with applications in a wide range of fields. The elbow method is then used to find the optimal value of k, which is four. The study's findings correspond to the Bank Administration Institute's customer segmentation scheme, which includes Disengaged Skeptics, High Valued, Satisfied, Traditionalists, and Majority Middles [7].

The other research was entitled "Application of Data Mining Techniques for Effective Customer Relationship Management of Microfinance: The Case of Wisdom Microfinance." This paper describes a study of data mining applications in microfinance that aid in the development of a classification model that aids in the prediction of a new borrower's status (highly privileged, moderately privileged, or less privileged) during loan decision-making in the organization. Based on the borrowers' corpus data obtained from the WISDOM microfinance, a classification model is built. Essential preprocessing activities have been used to clean and prepare the Experiment. The preprocessed dataset was then used in experiments with the J48 decision tree classifier of the WEKA 3.7.0 software, with different attributes and parameter settings, to find the best model. To predict the new customer class label, the classification model with the highest accuracy (78.502%) and the fewest leaves and tree size is built (highly privileged, moderately privileged, or less privileged) [9].

On the other hand, the research entitled “Bank Customer Churn Prediction: The Case of Commercial Bank of Ethiopia”. In this study, an effort is made to anticipate customer churn using machine learning algorithms. The work of cleaning and preparing the data for testing is finished after conducting business and data understanding. The experiment and model creation make use of machine learning algorithms like Support Vector Machine, K Nearest Neighbor, Nave Bayes, and Logistic Regression. R Studio and R programming were used to conduct each experiment. 48,051 data sets with 15 properties are used by the researcher. The data set includes the demographic and financial aspects of the customers. The accuracy of the KNN classifier was 99.91%, that of the SVM classifier was 92.4%, that of the Logistic Regression model was 93.8%, and that of the Nave Bayes classifier was 83.8%, according to the research results. In order to construct a bank customer churn prediction model for the Commercial Bank of Ethiopia, the KNN classifier is suggested [8].

With the researcher's knowledge in mind, there are no local studies on predicting customer lifetime value specific to the banking industry. Global studies have evaluated Data mining techniques in various fields, including banking for customer lifetime value analysis. However, most local studies focus on churn prediction, customer relationship management, and customer segmentation within the banking sector. A model for predicting customer lifetime value to identify valuable customers has not been researched locally yet. This gap limits Ethiopian banks' ability to measure customer value effectively, potentially leading to missed opportunities for targeted marketing, improved product offerings, and stronger customer relationships. This research aims to fill this gap by presenting a customer lifetime value prediction model using historical data from the Commercial Bank of Ethiopia. Other factors may influence predicting customers’ value, like demographic information and transactional data. Considering these factors can help CBE develop targeted marketing campaigns that are more likely to succeed. Predicting customer lifetime value (CLV) is vital for businesses to recognize and retain their most valuable customers. By analyzing CBE customer data and incorporating insights from previous research on relevant features used in other contexts, the study aims to pinpoint the most important attributes for predicting CLV within the CBE context. Recognizing these essential characteristics is vital for creating an effective CLV prediction model tailored to CBE's customers.

Machine learning algorithms vary in strengths and weaknesses for data management and prediction. The research aims to identify the best algorithm for creating a precise and dependable CLV prediction model by analyzing and comparing suitable algorithms, considering CBE's customer data peculiarities. Existing studies often use data from different countries and industries, highlighting a need for research comparing machine learning algorithms' performance in CLV prediction specifically at CBE. This comparison could assist CBE in selecting the algorithm that best fits their requirements.

CHAPTER THREE

METHODOLOGY

3.1 Overview

The research methodology should align with the research problem. It includes the approaches and perspectives of the entire research process. This chapter details the necessary methodologies for completing the thesis. Here, we explain the study technique, covering research strategy, approach, data collection methods, and data sources used to achieve the thesis's objectives. To create the proposed model and achieve customer lifetime value prediction goals by addressing research questions, this section also discusses tools and techniques used by the researcher.

3.2 Research Methodology

The study used the design science research methodology (DSRM). Design science is a problem-solving perspective that leverages engineering and science to create artifacts. This research utilizes the CRISP-DM framework, which demonstrates how data mining (DM) and Design Science Research (DSR) interact. The framework provides a systematic approach to data mining projects, ensuring each stage aligns with DSR's core principles. Essentially, DM techniques are vital for DSR, giving researchers skills to detect patterns in data, develop solutions, and evaluate their effectiveness in solving real-world problems.

The environment, as illustrated in the theoretical account below, defines the problem space in which the phenomena of interest exist. It includes various units such as people, organizations, and technologies that are found in existing technology, infrastructure, applications, communication architectures, and development capabilities. By combining these units, the researcher defines the business need or problem.

The knowledge base provides the raw materials from which the research is conducted. The foundations and methodologies make up the knowledge base. Prior design science research and results from reference disciplines provide the foundational theories, frameworks, instruments, constructs, models, methods, and instantiations used in the research studies development/build phase. Methodologies are in charge of the guidelines used in justifying and evaluating stages. Rigor is managed by applying existing foundations and methodologies correctly. Computational

and mathematical methods are primarily used in design science to evaluate the quality and effectiveness of artifacts [37].

Based on the concepts of design science research methodology Figure 3.1 depicts the conceptual research framework used to understand, execute, and evaluate this customer lifetime value prediction model. The design science research method is used to create the artifact, which is then assessed and evaluated through experimentation.

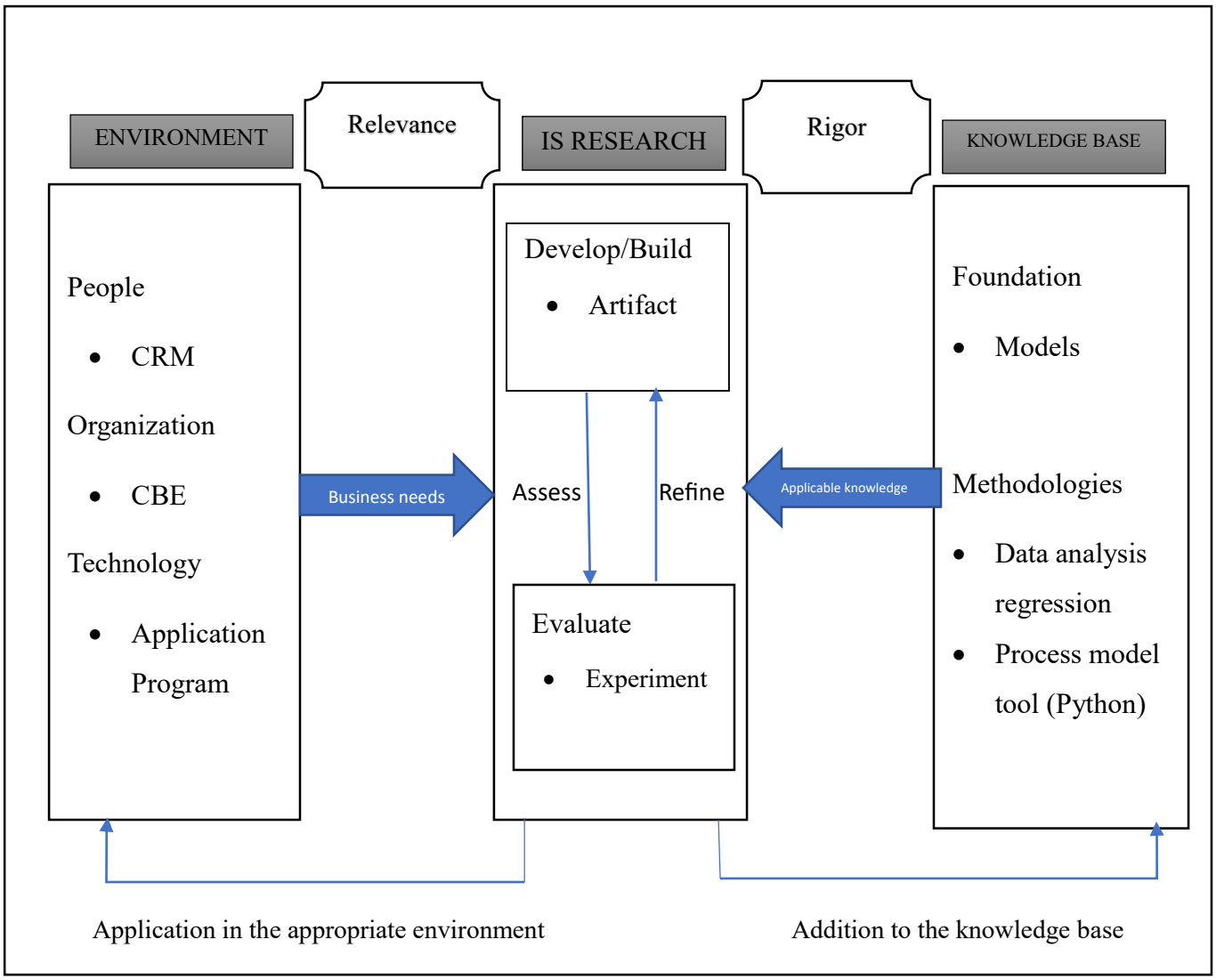


Figure 3.1 Design Science Research Methodology Adopted from [37]

The environment frameworks for the problem universe in which the events exist are depicted in the above figure. In this case, organizations are institutes that provide banking services which is

CBE, people are individuals with expertise in the field, and technology is the platform on which the application program runs. Based on these parameters, the researcher created a model using the aforementioned methodology and the Python tool and then tested the model to determine its validity.

In design science research methodology, we first understand the problem we are trying to solve: customer lifetime value prediction. Next, we identify the relevant variables for the model. Then, we collect the necessary data to train the model. After that, we develop and evaluate the model using real CBE customer data. Finally, based on the evaluation results, we refine the model. We consider the customer lifetime value prediction model as an artifact [37]. Design science research relies on DM methodologies, which offer tools and strategies for creating and evaluating artifacts that address issues through data analysis and knowledge extraction. The steps are perfectly aligning with the Cross-Industry Standard Process for Data Mining (CRISP-DM). The researchers ensure a systematic and well-documented approach to creating the CLV prediction model (artifact) using the DSR technique by following the CRISP-DM framework. We chose the design science research methodology for CLV prediction that allows us to develop real-world solutions. Design science research methodology is a good fit because it's a problem-solving method. CLV prediction is a problem-solving task as our goal is to create a model that predicts customer lifetime value. DSRM provides a systematic approach to developing solutions for problems. Additionally, we chose DSRM because it's an established methodology that has been effectively applied in various domains, giving us confidence that it would be an excellent fit for CLV prediction.

3.3 Data Mining Process Model

The Cross-Industry Standard Process for Data Mining (CRISP-DM) approach is used to achieve the stated objectives of this study. CRISP-DM is the most widely used framework for carrying out data science projects [25]. It is chosen for this study because it is a cross-industry standard methodology that can be implemented in any data-mining project, regardless of domain or destination. Furthermore, it is non-proprietary, technology-neutral, and has been widely used by field researchers for more than a decade. Similarly, the Design Science Process emphasizes a systematic approach to designing and creating solutions to address specific problems or needs.

The Design Science Process aligns well with CRISP-DM. The study essentially follows the fundamentals of the Design Science Process by using CRISP-DM, which provides an organized and iterative method for achieving research goals. The next section will cover in detail the six individual stages of CRISP-DM and how they were applied in the study.

3.3.1 Business Understanding

The Business understanding phase is critical to the success of any project because it ensures an understanding of objectives and needs. It's like creating the foundation for a house: Without a solid business understanding, the entire endeavor could fail. A thorough investigation of the banking industry is performed to achieve this critical information. This includes reviewing past research on customer relationship management, customer lifetime value, data mining process models, and data mining technologies. By going into these areas, new insights are discovered that provide chances for additional study and analysis.

To better understand the business domain of the Commercial Bank of Ethiopia and identify data mining problems, we have used various techniques. These include:

- ✓ Informal discussions with CBE senior management were conducted, as well as domain expert consultation, to gain a basic understanding of the problem.
- ✓ Visit the bank's official website and the internal portal.
- ✓ Investigating the bank's various printed media (different reports, magazines) as well as reviewing different literature.

According to the informal discussions with senior management, the bank is trying to attract and retain more customers while expanding the market to reach potential customers. Based on customer deposits, the bank determines high-value customers. This method aims to find customers who are more profitable for the bank. Traditional ways of figuring out customer lifetime value, however, might take a while and might not account for all important factors. To address this issue, we aim to build a model of customer lifetime value prediction by using data mining techniques like regression algorithms to datasets of CBE customers' transactions and demographic information. CLV is a powerful tool that can help CBE identify and target high-value customers. By analyzing data such as transaction history, frequency of transactions, and customer retention rates, CLV provides insights into each customer's long-term profitability. This

information can then be used to segment customers based on their potential value to the company. Implementing CLV strategies will enable the company to focus its marketing efforts on acquiring and retaining customers likely to generate higher revenue over their lifetime.

This describes the Business Understanding stage of the Design Science Process (DSP) in relation to CRISP-DM. Here's how it connects to the DSP: Identify the issue and specify objectives. The goal is explained as creating a customer lifetime value (CLV) prediction model based on data mining. Emphasize understanding the issue area and stakeholder demands according to approaches used in this phase, like document reviews and conversations by the DSP.

3.3.2 Data Understanding

After understanding the business, the next step is to analyze the available data. The Data Understanding phase builds on Business Understanding and focuses on identifying, collecting, and analyzing data sets for research. The goal is to thoroughly understand the available data. This procedure starts by searching a large database: The Commercial Bank of Ethiopia enterprise data warehouse. The exploration provides information on available data types and their relationships. Currently, the bank uses the T24 CORE-Banking system with an Oracle database for its business processes. Additionally, the bank has an enterprise data warehouse (EDW) that centralizes information from multiple sources and applications for analytics throughout the organization. This is done using an ETL tool to extract, transform, and load data from different databases into one place. The EFICAZ and ODI tools are used to extract data from various sources into the data warehouse. TOAD is then used as a graphical user interface tool to extract and convert this data into Excel format.

For this study, we extracted demographic and one year of transactional data of CBE’s customers from the EDW. We queried the EDW database with SQL during the data-understanding phase to identify potential data. Analyzing large datasets takes time and requires computational power. Using only one year of data can help reduce study costs and complexity. The TOAD tool session on a local computer is limited to one hour, so extracting data from large tables cannot be completed within that timeframe. To address this issue, we use server computers that are restarted three days a week to improve extraction and loading performance. However, due to these restarts, it's not possible to extract more than one year of transactional data.

The dataset used in this study is real customer data obtained from the Commercial Bank of Ethiopia's information management data warehouse and business intelligence department. The dataset includes 100,096 customer records with 19 attributes. The bank never exposes its customers' personal information, so fields like Account Number, Customer Name, House Number, Address, Telephone Number, Customer Mother Name, and other sensitive information are filtered. The customer data includes demographic information such as customer ID, date of birth, gender, and marital status; transactional data including credit transaction amount, debit transaction amount, and a total number of transactions; usage-related data like mobile banking usage status and card banking user status; and account-related data such as account opening date, working balances, account category, ownership description, and account opening district are availed for CLV prediction. Since the database contains a large amount of information, the attributes were identified based on a literature review to address the research question.

The total number of customers is more than 39.9 million. From this customer's data, the researcher sequentially selected 100,096 customer transaction records by joining different tables. As research suggests in customer lifetime value prediction modeling, we should consider customer demographic and transactional data. In our case, we joined the transactional data table, account-related data table, customer demographic data table, and digital channel users' status table to obtain the dataset with attributes. From these datasets, we extracted a total of 18 features that are most relevant to the CLV problem and best describe customers. By joining the above four tables with customer code we have extracted 100096 data sets with 18 attributes from the data warehouse.

Account data encompasses all information about a customer account. Account number (masked), LCY_CLOSING_BALANCE, account holders, account category, ownership, account activity account history. and the sector of the customer's account is the data field. The data fields define the customer's account.

The customer table contains information about the customer's facts, such as date of birth, gender, marital status, nationality, and the district in which the customer opened the account. The customers' table contains mostly static data. This means that the majority of the attributes contain customer information that is not supposed to change over time. Four attributes are nominated from this group: age, gender, marital status, and district because the bank does not expose the customer's privacy in any way. As a result, sensitive information such as the customer's name

(individual, group, or company), phone numbers, account numbers, and other identifying information are excluded from the study's data.

Transactional data is information captured during transactions. As a result, for this study, we have access to the customer's number of transactions, the number of credit transactions, the number of debit transactions, and the last transaction date.

Usage Data is data that is collected about how a product or service is used. This data can be used to improve the product or service, as well as to better understand customer needs and preferences. Such data can help businesses better understand their customer's usage habits. Commercial Bank of Ethiopia (CBE) offers various services supported by digital platforms, including mobile banking, internet banking, and ATM card banking. The attributes in each service show whether the customer is a user of the services or not.

This phase discusses applying DSP Data Understanding within CRISP-DM. It focuses on acquiring data from a bank's Enterprise Data Warehouse to extract customer information for creating a CLV prediction model. The phase highlights evaluating data quality limitations, like using only one year of transactions due to computational constraints. It emphasizes exploring available data sources aligned with study objectives. The researcher ensures thorough understanding by specifying gathering methods and feature analysis for effectively developing the prediction model.

3.3.3 Data Preparation

In this phase, we prepared the final data set(s) for modeling. The data preparation phase involved selecting, cleaning, integrating, and formatting data. During the data cleaning process, we removed any incorrect, inaccurate, improperly formatted, or useless information. The CBE data had inconsistencies in names, errors, and irrelevant details. The data cleansing step involves cleaning up data that has been affected by incorrect capitalization. Inconsistencies were found in the Gender attribute, with various mistakes in spelling for entries labeled as male (M, MA, Mael...) and female (F, Feamale, Female, FM...). To address all these issues, we used Python for the cleaning operation. Since Python is case-sensitive which means it handles capital and lowercase letters differently, we have to remove such inconsistencies in the dataset. To use the column inputs later on, we need to convert all text columns that have object data types to

numeric values. We use a label encoder to encode categorical variables as numerical values. We have also checked for missing values in the data set. Fortunately, the dataset does not have any missing values.

In the data preparation phase, we also calculated the estimated CLV for each customer using the formula CLV = Average Transaction Amount * Transaction Frequency *Average Customer Life Span. This variable is the target variable that we want to predict. Moreover, we had to check for the outliers in the dataset and remove it, since it is influential.

These explains how the Design Science Process (DSP) Data Preparation phase is linked to CRISP-DM. It details data cleaning, where errors are removed from customer data to ensure accuracy for model creation, such as a CLV prediction model. Additionally, it discusses data transformation by computing estimated CLV and converting numerical values into categorical variables to align with DSP's focus on artifact development. These steps highlight the researchers' preparation of data for modeling in line with DSP's goal of producing high-quality models through proper data handling techniques.

3.3.4 Modeling

Once the features have been selected, the outcome variables and ground truth have been labeled, and the data has been appropriately sampled and structured for analysis, the next step is to begin modeling and analyzing the dataset, and validating the resulting models. Modeling is frequently regarded as the most exciting work in data science. During this phase, we create and evaluate various models, frequently employing multiple modeling techniques. This study chooses supervised machine learning algorithms (linear regression, decision tree regression, and random forest) to investigate the analysis of customer lifetime value prediction of CBE customers. These modeling techniques were chosen because they are widely used and effective for predicting continuous variables such as customer lifetime value. Linear regression is a straightforward and easy-to-understand approach that aims to discover a relationship between predictor factors and target variables. Decision trees can capture non-linear correlations and interactions between variables, making them appropriate for large datasets. Random forest is an ensemble model that combines many decision trees to increase forecast accuracy and manage large datasets.

We split the dataset into training and testing sets to design the experiments. The training set was used to develop and train the models, while the testing set was used to evaluate their performance. The dataset was randomly divided into two parts: a training set with 70% of the data and a testing set with 30% of the data. This guarantees that the models are evaluated on previously unseen data to determine their generalizability. The reason why a 70:30 split is used is it is a good balance between having enough data for training and having enough data for testing. It allows the model to learn the underlying patterns in the data without overfitting the training data. It provides a realistic estimate of the model's performance on unseen data.

From the selected and extracted data from the data warehouse Feature selection is done in this phase using the recursive feature elimination (RFE) technique and chi-square test technique.

Recursive Feature Elimination (RFE) provides an appealing option. RFE progressively removes less significant features, resulting in a subset with the highest prediction accuracy. RFE assesses the influence of each feature on model performance using a machine learning method and an importance-ranking score. Recursive Feature Elimination is a feature selection approach used to determine the important features of a dataset. The procedure entails creating a model using the remaining characteristics after deleting the least significant elements continuously until the necessary number of features is attained. Although RFE can be employed with any supervised learning method [38].

The chi-square test aids in feature selection by examining the relationship between the features. It is used to determine the independence of two events. When the two attributes are independent, the observed count is close to the expected count, and hence the Chi-Square value is less. A high Chi-Square value suggests that the independence hypothesis is false. Simply said, the higher the Chi-Square value, the more dependent the feature is on the response and can be chosen for model training[39].

We may be more certain that we are choosing the most important features for the machine learning model by combining the two strategies. RFE is a wrapper approach that assesses how well a machine learning model performs on various subsets of features. A statistical technique known as the chi-square test can be used to identify characteristics that have a statistically significant correlation with the target variable. By combining two of these methods, we can choose a subset of features that are important to the model's performance and have a statistically

significant correlation with the target variable. We determined crucial factors that help predict CBE customer lifetime value using the aforementioned techniques.

The above section discusses how the Design Science Process (DSP) is integrated with the CRISP-DM model. It outlines the use of supervised machine learning algorithms for forecasting customer lifetime value (CLV) and emphasizes artifact development for CLV prediction models. The section explains data division for testing and training to ensure well-generalizing solutions, highlights feature selection methods to enhance data quality, and showcases researchers' adherence to DSP principles in constructing and evaluating CLV prediction models.

3.3.5 Evaluation

The evaluation phase focuses on identifying the model that best suits the demands of the business and on the following steps. It is essential to carefully assess the model and the processes used to build it before moving forward with the model's final implementation to make sure it fits the business objectives. As a result, evaluation is carried out using the evaluation methods outlined in chapter two-evaluation technique. We evaluated the performance of the prediction model using different regression model evaluation techniques such as R^2 and RMSE. The Evaluation phase of building customer lifetime value prediction includes measuring and comparing the performance of various models, reviewing the process, and determining the best course of action based on the evaluation results.

As cited in [40] more than one error measure is frequently necessary to evaluate model performance. To evaluate the models' performance, we applied two performance metrics. To evaluate a regression model, we would use a combination of R^2 and RMSE. R^2 is to determine how well the model fits the data. The model's performance on specific data points would then be evaluated using RMSE. This would provide us with a complete understanding of the model's strengths and weaknesses. These metrics assess the model's accuracy as well as its level of fit to the data. They enable us to compare and understand how well the models predicted the customer's lifetime value.

This section explains how the Evaluation stage of the Design Science Process (DSP) connects with CRISP-DM. Model validation ensures that a model meets business requirements, like estimating customer lifetime value. By comparing different models using R-squared and Root

Mean Squared Error, researchers select the most effective solution following DSP's goal of verifying solutions against set objectives.

3.3.6 Deployment

The deployment of a customer lifetime value prediction model involves the coordination of vital resources, which demands a collaborative effort of people as well as coordination of organizational procedures and technology. The model developed in this research is not deployed. As a result, the model's output should be organized to aid in decision-making and the design of marketing plans.

3.4 Tools

This section describes the tools used to preprocess, model, and evaluate CBE customer data. Anaconda Navigator is used for this research to access a web-based interactive computing notebook environment called Jupyter Notebook 6.0.3. Python programming was used for this research for data processing, data analysis, feature engineering, feature importance determination, modeling, and outcomes evaluation. Python is defined by its creators as "*an interpreted, object-oriented, high-level programming language with dynamic semantics. Its high-level built-in data structures, combined with dynamic typing and dynamic binding, make it very attractive for use as a scripting or glue language to connect existing components*[41]." It is preferred due to its open-source nature, flexibility, and simplicity of learning.

In this study, a personal computer with an Intel(R) Core (TM) i3-6006U CPU @ 2.00GHz 1.99 GHz, 4.00 GB of installed memory (RAM), and a 64-bit Microsoft Windows 10 pro–operating system was used for doing all the experiments.

CHAPTER FOUR

EXPERIMENTATION AND DISCUSSION OF RESULTS

4.1 Introduction

This chapter gives an in-depth analysis of the experiment. It begins by introducing the proposed architecture, which would serve as the foundation for the entire experiment. This architecture incorporates many components and operations required for data preparation.

The preprocessing tasks are critical because it ensure that the data is in a format that is suitable for the experiment. Cleaning the data, removing any outliers or inconsistencies, and putting it into a standard form that can be easily fed into the model is the goal of this stage. The process consists of feature extraction and selection, which involves identifying and selecting relevant features for use in the model. Following the preprocessing stage, the paper discusses the major experiments that were carried out.

In addition to reporting the experiments, this chapter examines the model's interpretation and evaluation. The interpretation of the results includes understanding the data's behavior and patterns and drawing knowledgeable conclusions. Evaluation, on the other hand, includes assessing the model's performance against standard metrics such as R^2 and RMSE. Overall, this chapter provides a thorough explanation of the experiments performed and data analysis.

4.2 The Proposed Architecture

The research's proposed architecture is shown in Figure 4.1. The proposed framework demonstrates the procedures used in the prediction of customer lifetime value using data from CBE customers. We use the CBE Customers dataset to verify the CLV prediction model.

Since it is important for evaluating the performance of our model on unseen data we should divide the cleaned dataset into training and testing data: First, we utilize the training data to determine the model's parameters. Following that, we evaluate the predictions made by the model for all of the testing data points to what we observed and use this comparison to determine the model's accuracy. In supervised Machine Learning (ML) activities, we frequently divide our dataset into a training dataset and a testing dataset. The training dataset is used to train the model, while the testing dataset is used to measure its performance. Empirical studies demonstrate that

using 20-30% of the data for testing and the rest 70-80% for training yields the greatest outcomes [42]. Accordingly, After Business Understanding and Data Understanding, the proposed architecture goes through six main stages to clean up the appropriate attributes that already exist in datasets, Data splitting:70 percent of the total dataset was used to train the model, and 30 percent was used to test the model's performance. Regression was performed using three supervised ML techniques, including Random Forest, linear regression, and decision tree regression, to produce the trained model. The train model and test dataset were then used to generate the customer lifetime value prediction output.

Using a 70:30 ratios for the training and test sets tends to produce the greatest outcomes. This is because a larger training set (70%) helps the model acquire more information and patterns from the data, resulting in improved performance. A smaller test set (30%) gives enough data to evaluate the model's performance.

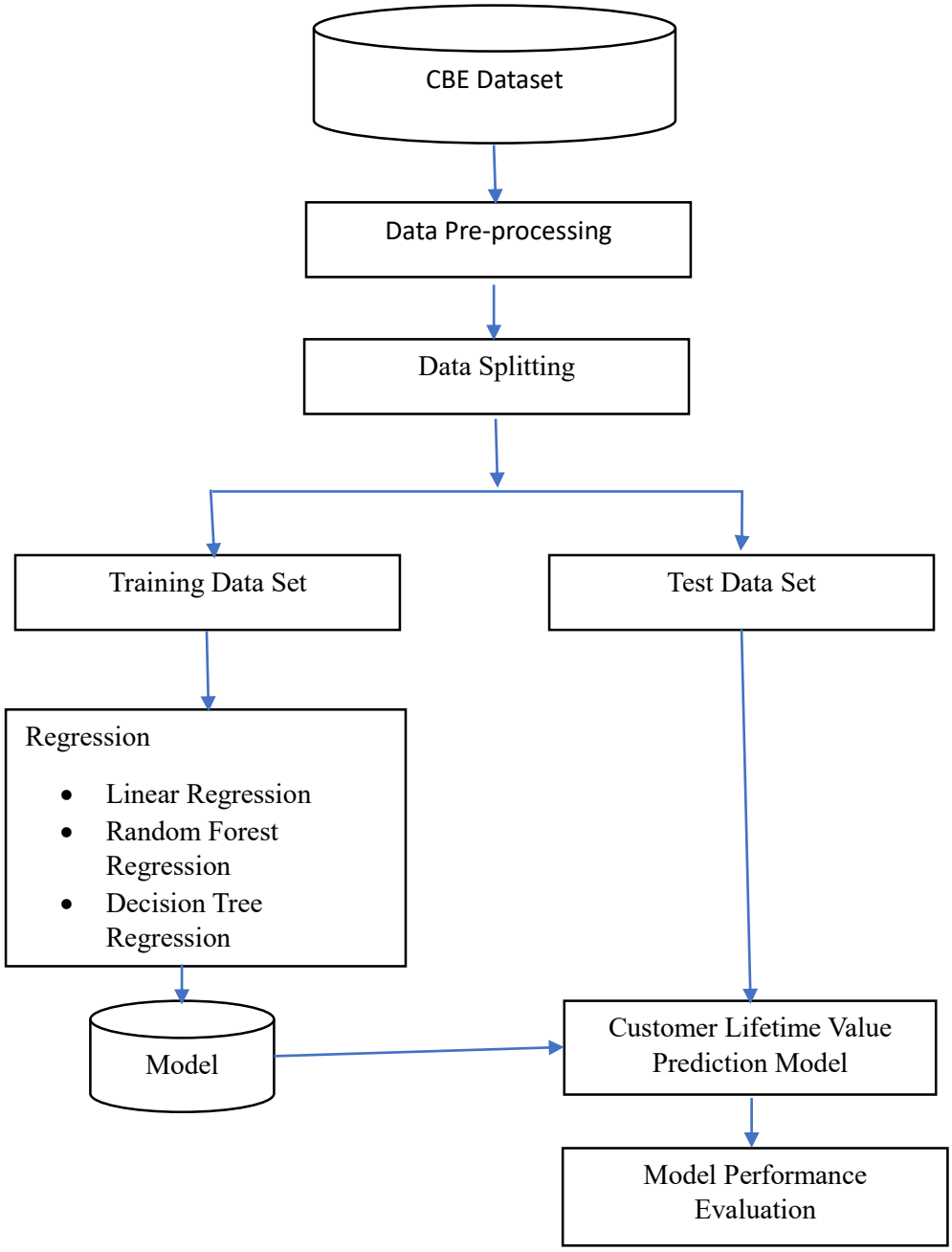


Figure 4.1 Proposed Architecture for Bank Customer Lifetime Value Prediction

4.3 Data Preprocessing

As stated in the previous chapter, the dataset used in this study is customer-related data obtained from the Information Management data warehouse and business intelligence department of the Commercial Bank of Ethiopia. The dataset we have selected from the Data warehouse consists of

100,096 customer records with 18 attributes. The dataset attributes are listed below, along with their descriptions and data types.

According to [10] other factors could be considered when predicting customers' value, such as demographic information, spending habits, and transactional information. By considering these other factors, CBE could create more targeted marketing campaigns that are more likely to be successful.

Table 4.2 Attribute Description

No.	Attribute Name	Description	Data Type
1.	CUSTOMER_CODE	Unique identification number for customers.	Number
2.	GENDER	Provide data on the distribution of the two main categories, male and female.	Text
3.	MARITAL_STATUS	gives the marital status of the customer. There are many different marital situations, including single, married, widowed, divorced, separated, and, occasionally, registered partnerships.	Text
4.	AGE	Although the data contains no age information, the date of birth will allow driving the customer's age by using the date the data is generated as a reference.	Number
5.	DISTRICT NAME	An attribute that provides information about a specific location/district in which the account owner resides, selected from the 33 Commercial Bank of Ethiopia districts.	Text
6.	OPENING_DATE	Specifies the exact date the account was opened or created.	Date time
7.	LCY_CLOSING_BALANCE	Current balance of the account in this case balance of the account as at December 31 2022.	Number
8.	OWNERSHIP	Indicates who has access to the account and who receives the account's proceeds. The	Text

		ownership type addresses who and how money in an account is managed.	
9.	CATEGORY	This is a type of CBE account that is used to categorize and track different types of financial transactions.	Text
10	ATM_CARD_STATUS	Customers can use these cards to withdraw cash and transfer funds. The attribute shows whether the customer is a user of ATM card or not.	Text
11	MOBILE_BANKING_STATUS	Customers can use these services to access their bank accounts via smartphones or tablets. The attribute shows whether the customer is a user or not user of mobile banking service.	Text
12	INTERNET_BANKING_STATUS	Give users online access to their bank accounts and the ability to carry out various financial operations. It shows whether the customer is a user or not user of internet banking services.	Text
13	CREDIT AMOUNT	A total number of transactions performed by the customer in one year that increase the account balance.	Number
14	DEBIT AMOUNT	Total number of transactions the customer performs in one year, reducing the account balance.	Number
15	TOTAL_TXN	Total number of transactions performed by the customer in one year.	Number
16	NET_MONTHLY_IN	Provides net monthly income of the customer.	Number
17	CUST_ED	Customer educational level	Text
18	LAST_TXN_DATE	Recent date the customer debited or credited	Date time

19	CLV	Estimated value of each customer over their relationship with the company	Number

Data preparation refers to any of the operations performed on extracted data to make the data more acceptable for experimentation and thus enhance the entire data mining activity. To that purpose, significant preprocessing operations were completed at this stage. Data profiling, data cleaning, data transformation, and feature selection are examples of these. Although as indicated in chapter two based on the literature we have calculated CLV (the predictor variable) for each customer.

As indicated in Chapter Two CLV is calculated for each customer using the formula CLV = Average Transaction Amount * Transaction Frequency *Average Customer Life Span.

The estimated value of a customer indicates how much money the business can expect to make from that customer. If a customer's estimated value is high, it means that they are a valuable customer and that the business can benefit from targeting them. This is because the bank is likely to make a profit from these customers.

4.3.1 Exploratory Data Analysis

This involves exploring and analysing the data to determine its characteristics, such as the number of records, the number of features, the data types of the features, and the existence of any missing values and outliers. Univariate, Bivariate, and Multivariate studies were used to highlight relevant parts of the data for additional study. Some of EDA's highlights are highlighted below.

```
{'whiskers': [<matplotlib.lines.Line2D at 0x1df7b282a08>,
  <matplotlib.lines.Line2D at 0x1df7b299c08>],
 'caps': [<matplotlib.lines.Line2D at 0x1df7b299cc8>,
  <matplotlib.lines.Line2D at 0x1df7b299d48>],
 'boxes': [<matplotlib.lines.Line2D at 0x1df7b298448>],
 'medians': [<matplotlib.lines.Line2D at 0x1df7b2a1e88>],
 'fliers': [<matplotlib.lines.Line2D at 0x1df7b2a1f48>],
 'means': []}
```

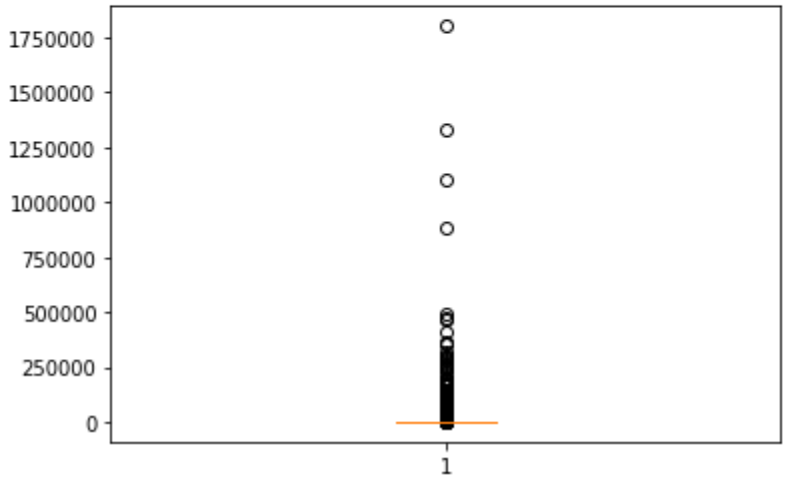


Figure 4.2 Box plot of CLV

The box plot reveals some significant outliers in CLV. We removed the outliers from the target variable (CLV), which removes the influential values.

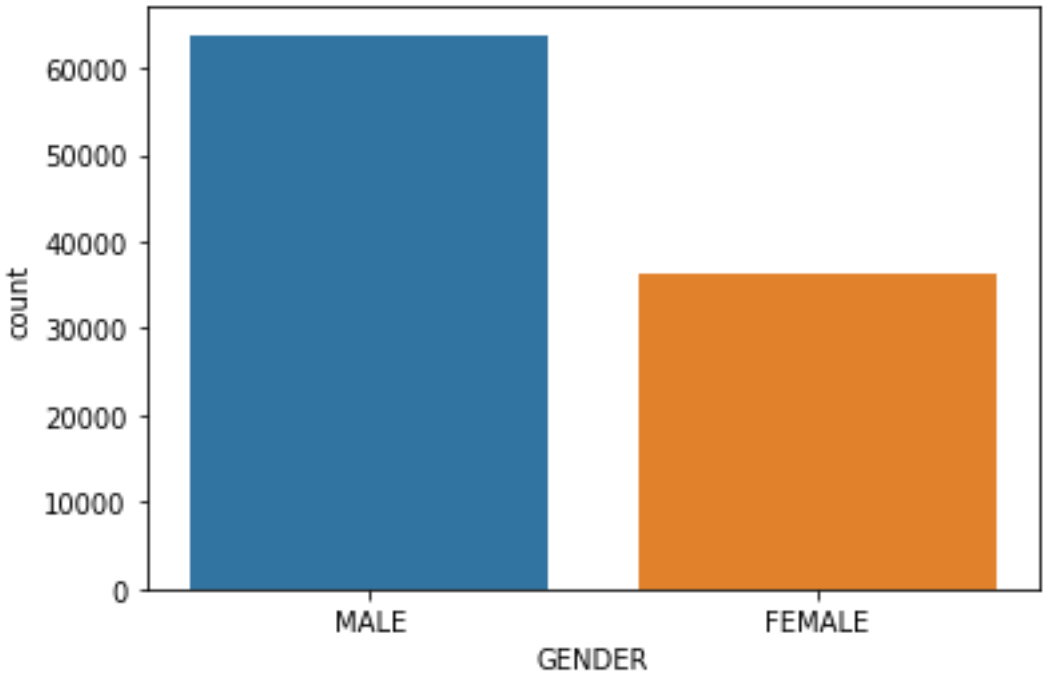


Figure 4.3 Gender Distribution of the Dataset

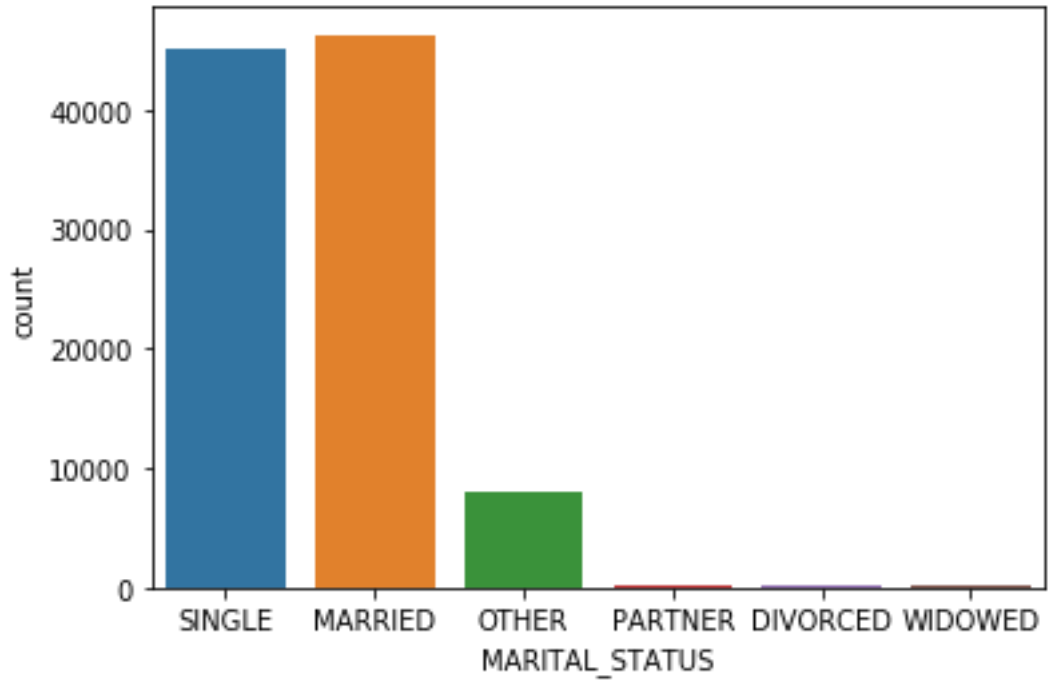


Figure 4.4 Marital Status Distribution of the Dataset

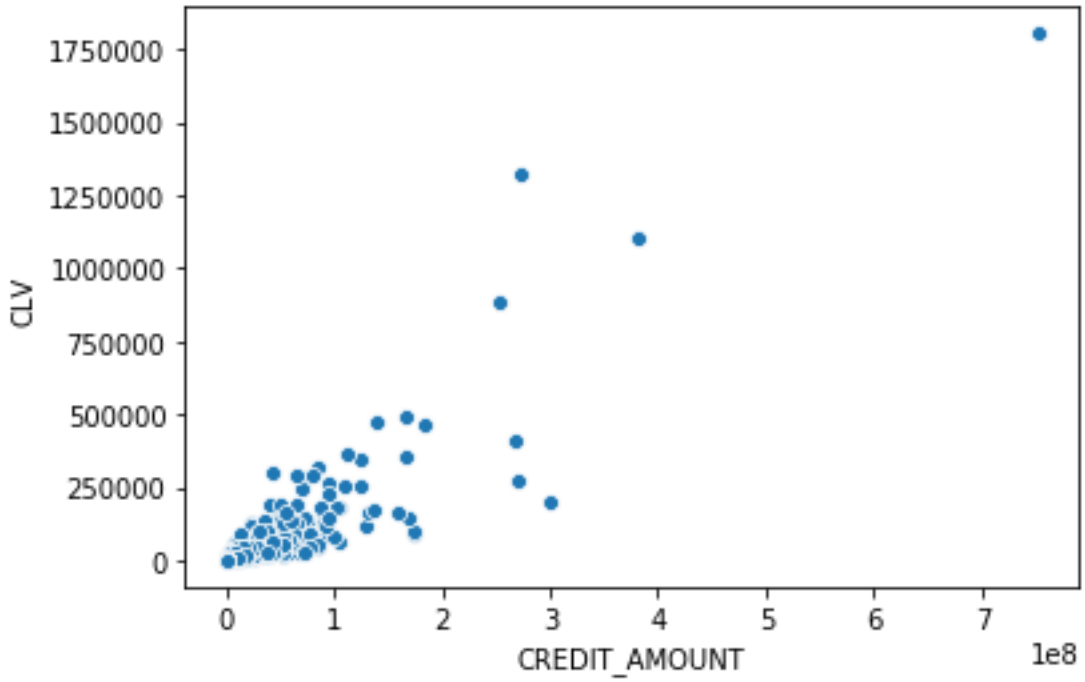


Figure 4.5 Gender Scatter Plot of Credit Amount and CLV

The above scatter plot shows the relationship between CLV and credit amount and it shows that as we have seen there is a correlation between these variables. The credit amount is the total

amount of deposit made by a customer while the CLV is the measure of the customer's creditworthiness. The scatter plot shows a positive correlation between the credit amount and CLV. This means that customers with higher credit amounts tend to have a higher value for the business.

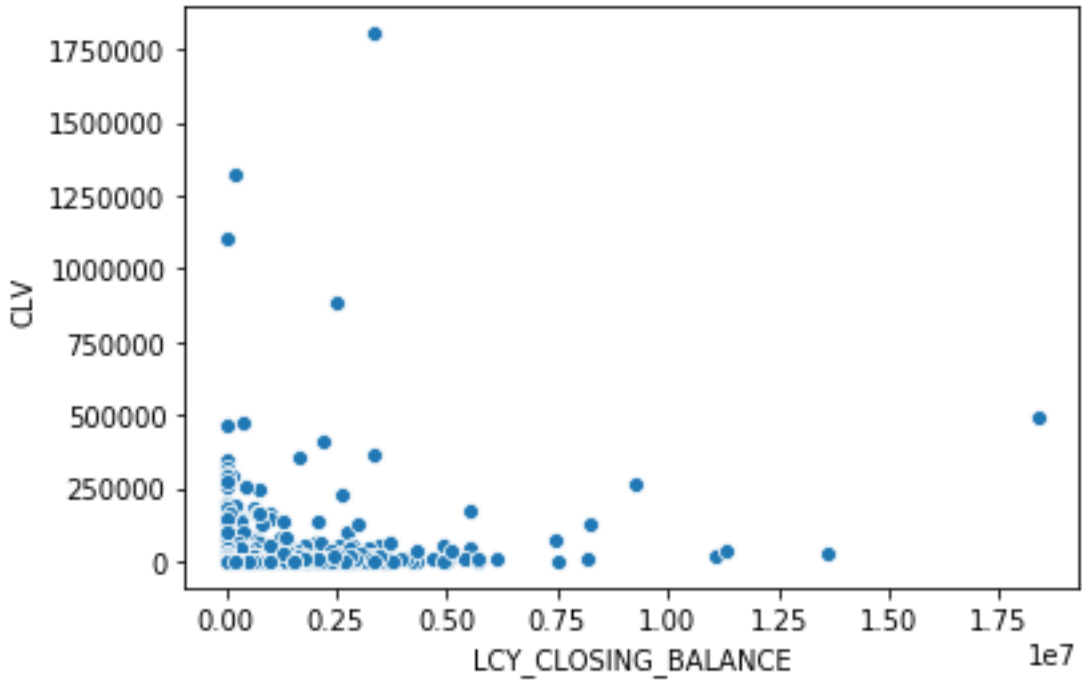


Figure 4.6 Scatter Plot of LCY_CLOSING_BALANCE and CLV

The scatter plot shows a correlation between the customer's lifetime value (CLV) and the working balance of the customer (LCY CLOSING BALANCE). CLV is a measurement of the total amount of money they are anticipated to spend with the bank. The scatter plot shows that customers with higher closing balances tend to have higher CLVs. This means that customers who have more money in their accounts are more likely to spend more money with the bank in the future. This is because these customers are more engaged with the bank and are more likely to continue using its products and services.

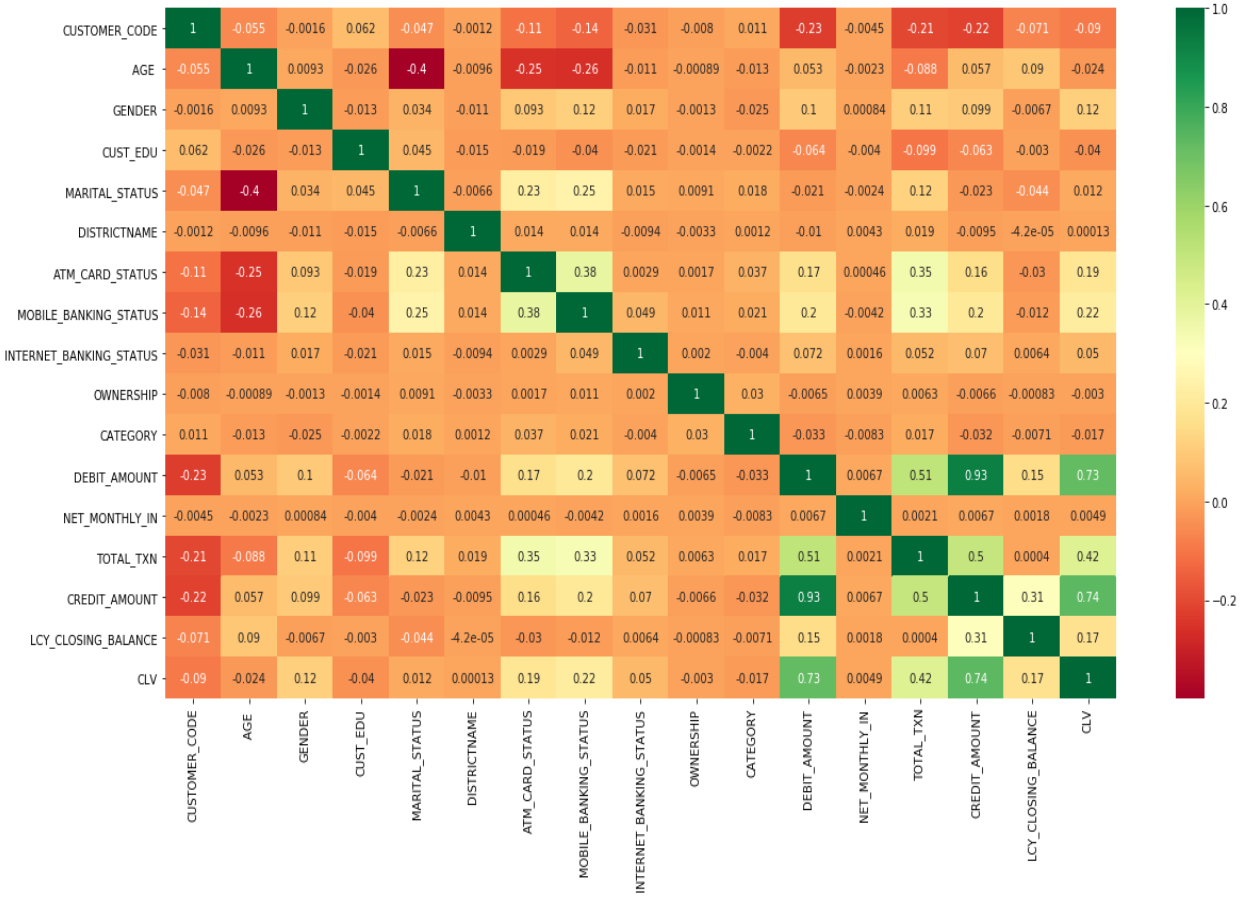


Figure 4.7 Heat map

A heat map is a visual representation of data that uses colors to represent different values. It is a tool for interpreting and easily understandable data visualization. It shows the correlation between different features (variables) in the dataset. The features are listed on the x-axis and the y-axis, and the correlation between each pair of features is represented by a color. The higher the number, the stronger the correlation.

Based on the description of the heat map, we can conclude that some of the features in the dataset are highly correlated with each other. For example, the CREDIT_AMOUNT, DEBIT_AMOUNT, and TOTAL_TXN features are strongly correlated with CLV. Other features, such as GENDER, CATEGORY, AND OWNERSHIP are less strongly correlated with CLV. Overall, the heat map is a useful tool for understanding the relationships between different

features in a dataset. It is possible to utilize it to find traits that are strongly associated with one another., as well as features that are negatively correlated with each other.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 100096 entries, 0 to 100095
Data columns (total 19 columns):
 #   Column                  Non-Null Count   Dtype
---  ------                  --------------   -----
 0   CUSTOMER_CODE           100096 non-null  int64
 1   AGE                     100096 non-null  int64
 2   GENDER                  100096 non-null  object
 3   CUST_EDU                100096 non-null  object
 4   MARITAL_STATUS          100096 non-null  object
 5   DISTRICTNAME            100096 non-null  object
 6   ATM_CARD_STATUS         100096 non-null  object
 7   MOBILE_BANKING_STATUS   100096 non-null  object
 8   INTERNET_BANKING_STATUS 100096 non-null  object
 9   OWNERSHIP               100096 non-null  object
 10  CATEGORY                100096 non-null  object
 11  DEBIT_AMOUNT            100096 non-null  int64
 12  OPENING_DATE            100096 non-null  object
 13  NET_MONTHLY_IN          100096 non-null  float64
 14  LAST_TXN_DATE           100096 non-null  object
 15  TOTAL_TXN               100096 non-null  int64
 16  CREDIT_AMOUNT           100096 non-null  int64
 17  LCY_CLOSING_BALANCE     100096 non-null  int64
 18  CLV                     100096 non-null  float64
dtypes: float64(2), int64(6), object(11)
memory usage: 14.5+ MB
```

As shown above there appears to be a lot of text/category information (Dtype 'object') and a few numerical columns (Dtypes 'int64' and 'float64'). The column 'CLV' is the one we want to predict. The data set contains 100,096 rows and 19 selected columns in the data preparation.

4.3.2 Data Cleaning

Finding and fixing any mistakes or inconsistencies in the data is required in the data cleaning step. This may include tasks such as handling missing values and correcting typos. Data cleaning is performed on the dataset with selected attributes which is the practice of removing inconsistencies and errors from data. Data cleansing is an essential stage in the data preparation process since it assures that the data is correct and complete. This procedure includes cleaning up data that has been affected by unusual names, errors, or wrong capitalization. The Gender attribute has inconsistencies, such as male (M, MA, Mael...) and female (F, Feamale, Female,

FM...) entries with different mistakes in spelling and there is inconsistency in the data set. Using Python, we have cleaned the inconsistency in gender attributes. We have also checked for missing values in the selected dataset. This dataset appears to have no missing values.

```
In [5]: #Checking Null values
        df.isnull().sum()
```

```
Out[5]: CUSTOMER_CODE              0
        AGE                        0
        GENDER                     0
        CUST_EDU                   0
        MARITAL_STATUS             0
        DISTRICTNAME               0
        ATM_CARD_STATUS            0
        MOBILE_BANKING_STATUS      0
        INTERNET_BANKING_STATUS    0
        OWNERSHIP                  0
        CATEGORY                   0
        DEBIT_AMOUNT               0
        OPENING_DATE               0
        NET_MONTHLY_IN             0
        LAST_TXN_DATE              0
        TOTAL_TXN                  0
        CREDIT_AMOUNT              0
        LCY_CLOSING_BALANCE        0
        CLV                        0
        dtype: int64
```

4.3.3 Data Transformation

This involves transforming the data into a format appropriate for the chosen modelling algorithm. This could involve activities like categorical feature encoding and feature creation from already existing ones. We have created an age attribute from the date of birth feature by subtracting the date of birth from the current date. Then categorize into three group since it has many values.

In machine learning, data representation refers to how data is organized and arranged before training a machine learning model. This involves converting raw input data into a format that the learning algorithm can process. The most common type of data representation is numerical, where data is transformed into numerical values. To use the column inputs later on, we need to convert all text columns to numeric values. We use a label encoder to encode categorical variables as numerical values. The label encoder assigns a unique integer to each category.

4.4 Feature Selection

First, data reduction must be conducted by removing unneeded or less significant features from the original data. This is done by the study's objectives. Since CUSTOMER_CODE features are attributes that are used to identify the customer uniquely there is no impact on the study. As a result, these attributes were deleted to compress the data to only the most significant ones; this reduces the work necessary for subsequent processing. The rest are retained for further preprocessing. (A snapshot of the dataset used is shown in Appendix A).

Feature selection involves identifying and selecting the subset of features that are most important for the modelling algorithm to learn from. This can help to improve the performance of the model and reduce the risk of overfitting. Numerous elements influence the success of data mining algorithms on the task at hand. One such factor is the data's quality. Knowledge discovery during training becomes more difficult when information is useless or redundant, or when data is noisy and inaccurate. The practice of detecting and deleting as much irrelevant and redundant information as feasible is known as attribute subset selection. The importance that learning algorithms place on attribute selection varies [43].

We used the chi-square test and RFE approaches to select features. The chi-square test is suitable for feature selection with categorical data as it checks independence between categorical variables. RFE helps identify key predictive traits by iteratively eliminating variables based on their relevance. By combining these strategies, we reduce dataset size and enhance model performance by pinpointing crucial features.

For the experiment, we selected features that were chosen by both techniques. Selecting features that were chosen by both chi-square test and RFE is a good way to improve the accuracy of the machine learning model.

Selected important feature ranking using the RFE technique is shown below.

Out[244]:

	Importance	Columns
0	1	AGE
14	1	TOTAL_TXN
13	1	LAST_TXN_DATE
10	1	DEBIT_AMOUNT
4	1	DISTRICTNAME
15	1	CREDIT_AMOUNT
16	2	LCY_CLOSING_BALANCE
11	3	OPENING_DATE
5	4	ATM_CARD_STATUS
12	5	NET_MONTHLY_IN
9	6	CATEGORY
2	7	CUST_EDU
3	8	MARITAL_STATUS
7	9	INTERNET_BANKING_STATUS
6	10	MOBILE_BANKING_STATUS
8	11	OWNERSHIP
1	12	GENDER

Figure 4.8 RFE technique Feature Ranking

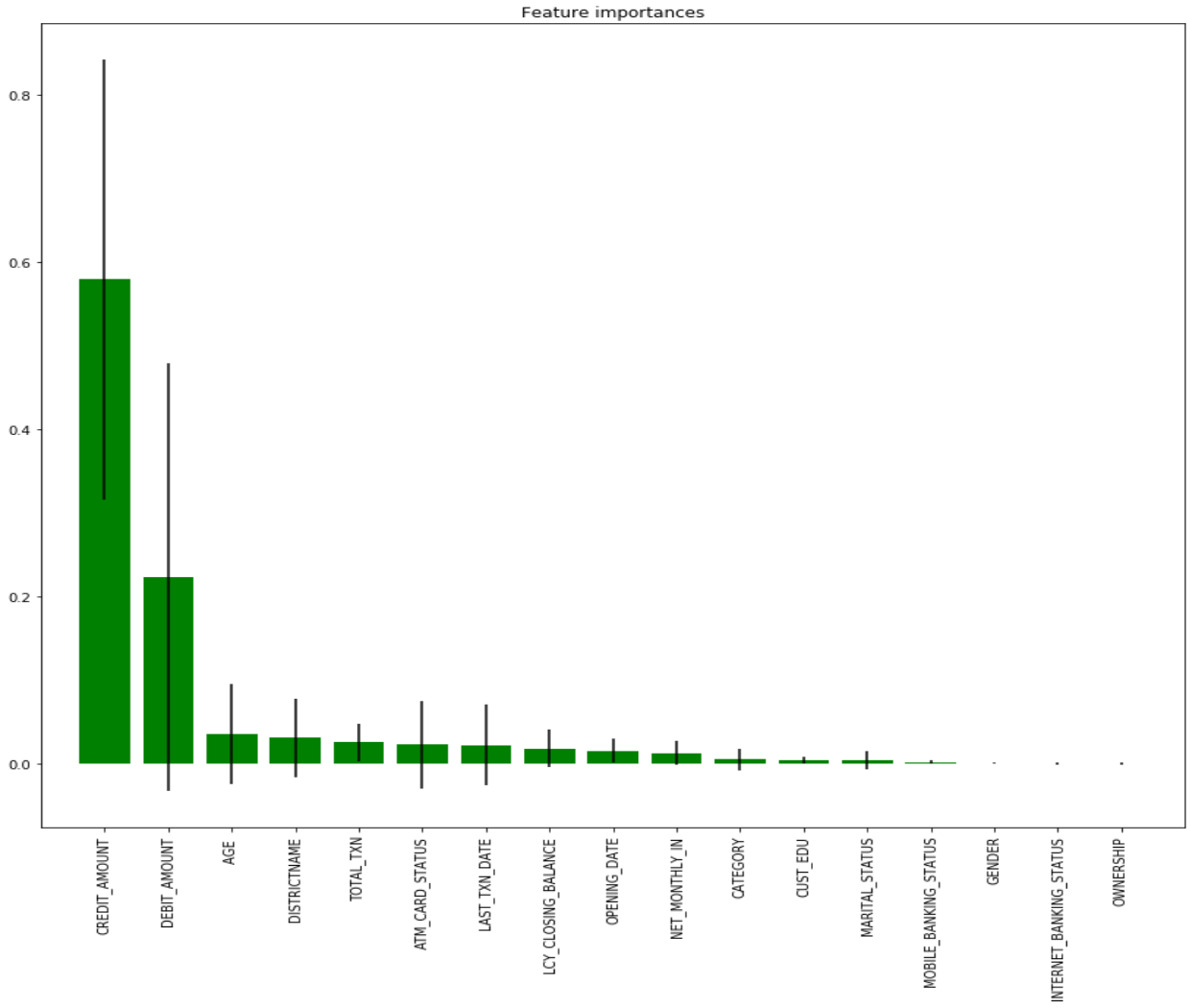


Figure 4.9 Feature Importance Ranking

The selected most important features using the chi-square test technique are shown below.

	Name of the column	Score
13	CREDIT_AMOUNT	7.543741e+09
10	DEBIT_AMOUNT	7.455296e+09
11	NET_MONTHLY_IN	6.463842e+09
14	LCY_CLOSING_BALANCE	9.962107e+08
12	TOTAL_TXN	2.234573e+06
0	AGE	3.805559e+03
5	ATM_CARD_STATUS	3.164738e+03
6	MOBILE_BANKING_STATUS	2.973465e+03
4	DISTRICTNAME	2.944789e+03
2	CUST_EDU	1.225214e+03

Figure 4.10 Feature Ranking Chi-Square Test Technique

4.5 Modeling using Linear Regression

The development of the linear regression prediction model is covered in this subsection. In this experiment, a linear regression approach is used to develop a prediction model.

4.5.1 Experiment One

Based on data collected from CBE, the dataset includes customer demographic information and transactional data. Both transactional and demographic characteristics were considered attributes. The dataset consists of 100,096 observations and 19 variables, with CLV as the dependent attribute and the remaining values as predictor (independent) attributes. After completing all preprocessing tasks and preparing our dataset for experimentation, we conducted experiments using all available variables. The preparation procedures included data cleansing, outlier removal, and categorical variables conversion to numerical variables.

4.5.1.1 Creating Training and Test Dataset

The y-column (the target) and the components (X), which are independent columns, must first be separated from the dataset. This is essential since our model will predict the y using the X columns. On the training dataset, we execute the linear regression algorithm, and we compare the outcomes to the test dataset to ensure accuracy. In this work, the dataset was divided into training and testing halves in a 70:30 ratios. Customer Lifetime Value (CLV), the target variable, is present in the training dataset but absent from the test dataset. This is crucial for assessing how well the model works with unobserved data. On the test set, train the model. In order for the model to understand how the characteristics relate to the CLV target variable, training data must be fed into it.

4.5.1.2 Evaluating Model Performance

The model's ability to match the data is shown by the (RMSE). It is calculated using the square root of the mean squared error (MSE), which is the average of the squared discrepancies between the expected and actual values. An accurate fit is suggested by a reduced RMSE.

The dependent variable's variance is best explained by the independent components, as indicated by the R-squared value. It is calculated by dividing the squares from the regression by the sum of all the squares. A better fit is indicated by a higher R-squared value.

In other words, R-squared indicates how much of the variance in the dependent variable is explained by the independent variables, but RMSE indicates how far the model's predictions are from the actual values.

```
In [92]: model.Linear_Regression_Model(df)
Out[92]: (58.042855528059725, 0.5632066438680678)
```

The RMSE (Root Mean Squared Error) of the regression model is 58.0428. This means that the model's predictions are, on average, 58.0428 units away from the actual values.

The R-squared of the regression model is 0.5632. This means that the independent variables explain 56.3% of the variation in the dependent variable.

In other words, the regression model is not very accurate, but it is still able to explain a significant portion of the variation in the dependent variable.

4.5.2 Experiment Two

The dataset has 100,096 rows and 9 columns. The CLV column is the target variable, and the other columns are predictor variables. After preprocessing the data, we used RFE and the chi-square test to select the most important features. We then removed all columns except for the selected features. Next, we split the data into training and test sets in a 70:30 ratios. We trained the model on the training set and evaluated its performance on the test set using regression model evaluation metrics.

```
In [50]: model.Linear_Regression_Model(df)
Out[50]: (57.41016358278667, 0.572706656027263 9)
```

As shown above the RMSE and R^2 values of the regression model is 57.4101 and 0.5727 respectively. The regression model's predictions are, on average, 57.41 units away from the actual values, and the independent variables explain 57.27% of the variation in the dependent

variable. In other words, the model is not perfect, but it is able to explain a significant portion of the variation in the dependent variable.

4.6 Modeling using Decision Tree Regression

This subsection deals with how the decision tree regression prediction model is developed. In this experiment, a Prediction model building is done using a decision tree regression. The dataset consists of 100,096 observations and 9 variables, where CLV is the dependent attribute and the rest values are predictor (independent) attributes. After performing all the preprocessing tasks and making ready our dataset for the experiment we experimented using all the variables we had. The pre-processing tasks includes cleaning the data, removing outliers, and converting categorical variables to numerical variables.

4.6.1 Experiment One

4.6.1.1 Creating Training and Test Dataset

Data splitting involves splitting the data into two sets: a training set and a test set. The training set is used to train the model, and the test set is used to evaluate the performance of the model on unseen data. To begin, we must divide the dataset into the y-column (the target) and the components (X), which are independent columns. This is required since we will be using the X columns to predict the y in the model. The linear regression algorithm is run on the training dataset, and the results are validated against the test dataset. In this study, we split the dataset in two with a 70:30 ratios for the training and test datasets. CLV (Customer Lifetime Value) is the target variable, which is present in the training but not in test datasets.

The R-squared number expresses how well the independent factors explain the variation in the dependent variable. It is calculated by dividing the regression sum of squares by the total sum of squares. A greater R-squared value denotes a better fit.

```
In [93]: model.Decision_Tree_Model(df)
Out[93]: (70.44291734656832, 0.3566413405992608)
```

The RMSE of the model is 70.4429 and the R-squared value is 0.3566. This means that the model does not fit the data very well, especially compared to a linear regression model. In

simpler terms, the model is not very accurate at predicting the target variable. This is supported by the high RMSE value and the low R-squared value.

4.6.2 Experiment Two

We started with a dataset of 100,096 rows and 19 columns. We cleaned the data, removed outliers, and converted categorical variables to numerical variables. Then, we used the RFE and chi-square test feature selection algorithms to identify the most important variables for predicting CLV. Finally, we trained a decision tree regression model on the selected variables. The selected attributes are described in the feature selection section of this chapter.

```
In [44]: model.Decision_Tree_Model(df)
Out[44]: (46.246610827758175, 0.7227264893131586)
```

The model explains 72% of the variation in the CLV, which is a significant portion. The RMSE of the model is 46.2466, which is slightly better than the performance of the first model that used all of the attributes.

The fact that the model performs slightly better than the first model that used all of the attributes suggests that the feature selection process was effective. This means that the selected features are more informative than the other features in the dataset.

4.7 Modeling using Random Forest Regression

This subsection describes how to develop a Random Forest Regression prediction model. In this experiment, we used a Random Forest Regression algorithm to build a prediction model for the customer lifetime value (CLV). The dataset had 100,096 rows and 19 columns, with CLV as the dependent variable and the other columns as predictor variables.

4.7.1 Experiment One

We cleaned the data, removed any outliers, and converted categorical variables to numerical variables before using all of the variables in the dataset to train our model.

4.7.1.1 Creating Training and Test Dataset

Once the data has been preprocessed and the necessary libraries have been imported, the dataset needs to be split into two parts: the target variable (y) and the predictor variables (X). This is necessary because the predictor variables will be used to predict the target variable in the model. The random forest regression algorithm is then trained on the training dataset and evaluated on the test dataset. In this experiment, the dataset was split into a 70:30 ratios for the training and test datasets, respectively. The target variable is CLV (Customer Lifetime Value), which is present in the training dataset but not in the test dataset.

```
In [94]: model.Random_Forest_Model(df)
Out[94]: (49.76217425405421, 0.6789466700186828)
```

The R-squared value is a measure of how well the independent variables (X) explain the variation in the dependent variable (y). It is calculated by dividing the sum of squares of the regression by the total sum of squares. A higher R-squared value indicates a better fit. In this case, the R-squared value is 0.6789, which indicates that the model explains 67.89% of the variation in the dependent variable. This is a good fit, and it suggests that the model can be used to make accurate predictions about the dependent variable.

The root mean squared error (RMSE) is a measure of how far off the model's predictions are from the actual values. A lower RMSE indicates a better fit. In this case, the RMSE is 49.76, which means that the model's predictions are, on average, 49.76 units away from the actual values.

4.7.2 Experiment Two

We started with a dataset of 100,096 rows and 19 columns. We cleaned the data to remove any errors or inconsistencies, removed outliers (data points that were significantly different from the rest of the data), and converted categorical variables to numerical variables. Next, we used two feature selection algorithms, called RFE and chi-square test, to identify the most important

variables for predicting CLV (Customer Lifetime Value). Feature selection is the process of identifying the most informative variables in a dataset. This can help to improve the accuracy of a model and reduce the risk of overfitting.

Once we had identified the most important variables, we trained a random forest regression model on those variables. A random forest regression model is a type of machine learning model that can be used to predict continuous values. It works by constructing a number of decision trees and then averaging their predictions.

```
In [46]: model.Random_Forest_Model(df)
Out[46]: (31.908079957710996, 0.8680073031268528)
```

The regression model has an RMSE of 31.9080 and an R-squared value of 0.8680. This indicates that the model is able to explain a large proportion of the variance in the data. The random forest model has the highest prediction accuracy, followed by the decision tree and linear regression models.

4.8 Comparison of Machine Learning Models

Model selector using k-fold cross-validation is a technique used to select the best machine-learning model for a particular problem. It works by dividing the training data into k folds, training each model on k-1 folds, and evaluating its performance on the remaining fold. This process is repeated k times, with each fold being used as the validation set once. The model with the best average performance on the validation sets is selected as the best model.

In the below screenshot, the three supervised machine learning algorithms were evaluated using k-fold cross-validation with k=5. The random forest regression algorithm had the highest average R-squared and lowest average RMSE on the validation sets, indicating that it is the best model for the prediction of customer lifetime value prediction at the CBE dataset.

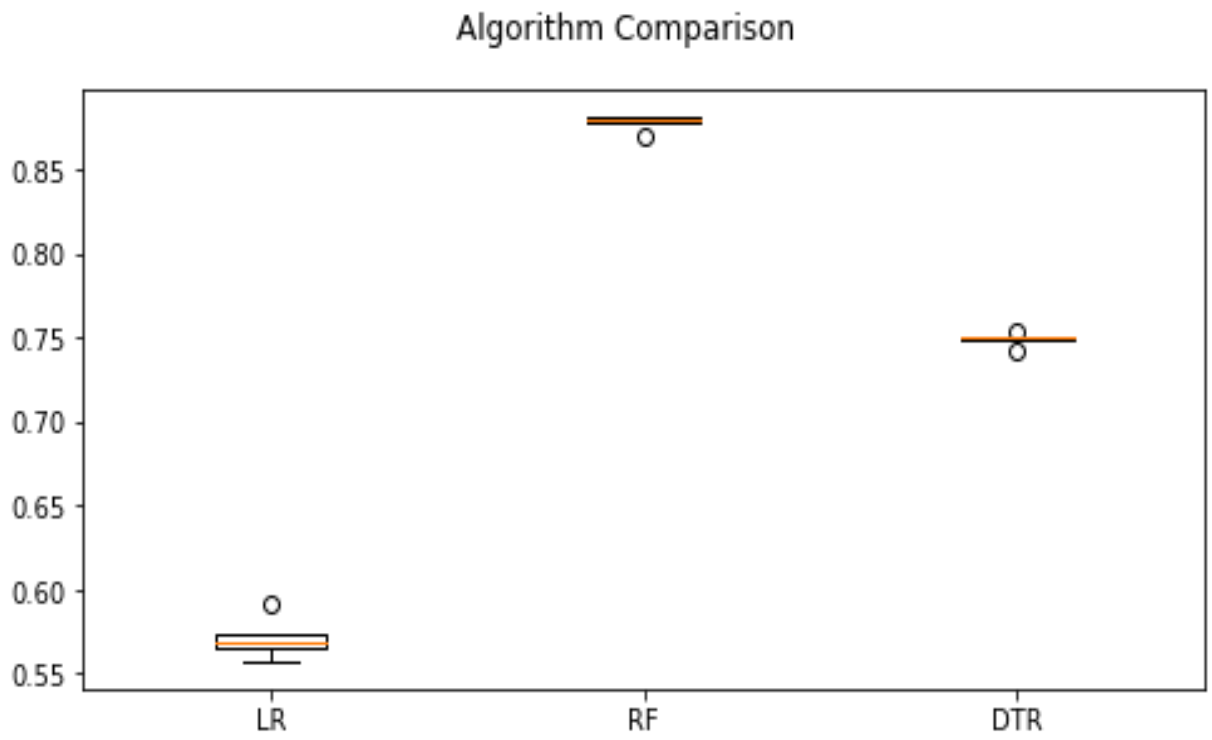


Figure 4.11 Comparisons of the algorithms

Based on the values shown in the image, the random forest model has the best performance on the validation set. This means that the random forest model is likely to make the most accurate predictions on new data.

Table 4.3 Regression Model Evaluation

	Before feature selection		After feature selection	
Model	RMSE	R^2	RMSE	R^2
Linear regression	58.0428	0.5632	57.4101	0.5727
Decision Tree Regression	70.4429	0.3566	46.2466	0.7227
Random Forest Regression	49.7621	0.6789	31.9080	0.8680

As clearly shown in the table the random forest regression model has the highest R^2 and the least RMSE, followed by the decision tree regression model and the linear regression model. This suggests that the random forest regression model is the best model for the CBE data set to predict customer lifetime value.

4.9 Result and discussion

The following section summarizes the main findings of this study. As was previously stated in chapter one, the primary goal of developing the prediction model is to assign a value to each customer in order to forecast the potential value of a new customer based on these characteristics. The dataset that was utilized for the experiment is thoroughly tested, which is detailed in the data preprocessing section. The CBE dataset contains 100,096 observations, with 30% allocated as a test sample and the remaining 70% designated for training. The prediction is verified using Linear Regression, Decision Tree Regression, and Random Forest Regression Machine Learning Algorithms. The results are then presented and discussed separately. The summary includes RMSE and R2 values in a table format to show the model's performance.

4.9.1 Results on Linear Regression Algorithm

A separate check was conducted using a Linear Regression model with all features extracted from CBE. Then the two feature selection techniques (chi square and RFE) discussed in the feature selection section were applied. We checked the performance of the model through RMSE and R2 values. The results of the experiment are presented below in a table summarizing two experiments (before feature selection and after feature selection) conducted in the Linear regression model.

Table 4.4 Results on Linear Regression Algorithm

	Before feature selection		After feature selection	
Model	RMSE	R^2	RMSE	R^2
Linear regression	58.0428	0.5632	57.4101	0.5727

The results indicate that after feature selection, the Root Mean Squared Error (RMSE) decreased from 58.0428 to 57.4101, while the R-squared (R^2) value increased from 0.5632 to 0.5727 for the Linear Regression model. This suggests that the model's predictive performance improved slightly after feature selection, as indicated by the lower RMSE and higher R^2 values. Overall, the model is better able to explain the variance in the data and make more accurate predictions with the selected features.

4.9.2 Results on Decision Tree Regression Algorithm

Like the experiment using Linear regression, we also conducted an experiment on the decision tree regression algorithm with 100,096 rows and 19 columns in the CBEs dataset. We evaluated the model performance through RMSE and R^2. The results of the experiment (before feature selection and after feature selection) are presented below.

Table 4.5 Results on Decision Tree Regression Algorithm

	Before feature selection		After feature selection	
Model	RMSE	R^2	RMSE	R^2
Decision Tree Regression	70.4429	0.3566	46.2466	0.7227

The results show a significant improvement in the performance of the Decision Tree Regression model after feature selection. The Root Mean Squared Error (RMSE) decreased from 70.4429 to 46.2466, while the R-squared (R^2) value increased from 0.3566 to 0.7227. This indicates that the model's predictive accuracy has greatly improved after selecting the features. The lower RMSE and higher R^2 values suggest that the model is now better at predicting the target variable and explaining the variance in the data.

4.9.3 Results on Random Forest Regression Algorithm

The experiment used Linear regression and decision tree regression, then conducted a new one using the random forest regression algorithm on the CBEs dataset with 100,096 rows and 19 columns. We evaluated the model performance using RMSE and R^2. Below is a table summarizing two experiments (before feature selection and after feature selection) done in the Random Forest Regression model.

Table 4.6 Results on Random Forest Regression Algorithm

	Before feature selection		After feature selection	
Model	RMSE	R^2	RMSE	R^2
Random Forest Regression	49.7621	0.6789	31.9080	0.8680

The above results demonstrate an improvement in the performance of the Random Forest Regression model after feature selection. The Root Mean Squared Error (RMSE) decreased from 49.7621 to 31.9080, while the R-squared (R^2) value increased from 0.6789 to 0.8680. This indicates that the model's predictive accuracy has enhanced after selecting the most important

features. The lower RMSE and higher R^2 values suggest that the model is now more adept at predicting the target variable (CLV) and explaining the variance in the data.

The methods for designing the three prediction models Linear Regression, Decision Tree Regression, and Random Forest Regression Machine Learning Algorithms techniques as well as the methods for performance evaluation were covered in this chapter. In the meantime, the regression models' RMSE and R^2 values were computed. After that the research questions raised have been answered as follows.

Research Question One: “What are the significant attributes that help to predict the lifetime value of CBE customers?” In CBE, CREDIT_AMOUNT, DEBIT_AMOUNT, AGE, DISTRICTNAME, TOTAL_TXN, ATM_CARD_STATUS, LCY_CLOSING_BALANCE, NET_MONTHLY_IN, and CUST_EDU are effective and relevant (important) predictors of customer lifetime value.

Research question Two:” Which algorithm best suits to customer lifetime value prediction modeling at CBE?” The Random Forest regression was shown to be the best fit for the customer lifetime value prediction model at the CBE dataset. The R^2 is 86.8%.

Research question three “To what extent the proposed model performs in customer Lifetime Value prediction?” Linear Regression performed at 57.27%, Decision Tree Regression at 72.27%, and Random Forest Regression at 86.80%. These models aimed to show how data mining can help in decision-making by identifying high-value customers and their main predictors.

CHAPTER FIVE

CONCLUSIONS AND RECOMMENDATIONS

5.1 Overview

The complete study results are described in this chapter. It is split into two sections: conclusions and recommendations for future works. The conclusions section provides a complete assessment of the study's important findings and outcomes. It summarizes the main findings and interpretations derived from the data analysis and investigation. It also provides a concise summary of the research objectives and how they were met. Moving on to recommendations, this section provides useful suggestions for future research in the field. It identifies potential areas for improvement and future study directions.

5.2 Conclusions

The data for the experiment was gathered from the commercial bank of Ethiopia's data warehouse and business intelligence department on Excel sheets, about 100,096. After performing the appropriate data preprocessing steps on the dataset, 100,096 data were ready for experimentation. Linear regression, random forest regression, and decision tree Regression algorithms were used to generate regression and prediction models. Python programming is used to simulate all of the experiments. Finally, Root Mean Square Error and R^2 values were calculated for each regression model.

The Root Mean Square Error for the linear regression model was high at 57.5259, indicating a considerable level of error in predictions. The R^2 score was 0.5692, suggesting that the linear regression model explained only 56.9% of the variation in the data.

When compared to the linear regression model, the Decision Tree Regression model fared better. It had a smaller Root Mean Square error of 45.8296, indicating fewer mistakes in predictions. Th e R^2 score is higher at 0.7266, suggesting that the Decision Tree Regression model explained 72. 66% of the variance in the data.

In conclusion, the random forest regression model outperformed the other two models mentioned above. it had the smallest root mean square error of 31.9782, suggesting the lowest degree of prediction error. The R^2 value is 0.8668 indicating that the random forest regression model

explained 86.6% of the variance in the data. In terms of both the two criteria, the Random Forest Regression model performed the best from decision tree and linear regression. This study is a good starting point for further research on predicting customer lifetime value.

Based on the aforementioned conclusion, both of the research questions have been addressed as follows.

- ✓ Research Question One: "What are the significant attributes that help to predict the lifetime value of CBE customers?" In CBE, CREDIT_AMOUNT, DEBIT_AMOUNT, AGE, DISTRICTNAME, TOTAL_TXN, ATM_CARD_STATUS, LCY_CLOSING_BALANCE, NET_MONTHLY_IN, and CUST_EDU are effective and relevant (important) predictors of customer lifetime value.
- ✓ Research question Two:" Which algorithm best suits to customer lifetime value prediction modeling at CBE?" The Random Forest regression was shown to be the best fit for the customer lifetime value prediction model at the CBE dataset. The R^2 is 86.8%.
- ✓ Research question three "To what extent the proposed model performs in customer Lifetime Value prediction?" Linear Regression by 57.27%, Decision Tree Regression by 72.27% and Random Forest Regression by 86.80 were modeled.

The studies' results indicate that data mining technologies can be effectively implemented on banks to build customer lifetime value predictions. As a result, banks in Ethiopia can use data mining technology to predict the lifetime value of customers. We assumed that by doing so, CBE would be able to identify and predict their customer's lifetime value and achieve the anticipated investment returns.

Modeling might have been more successful if the dataset was larger and there were more and more different types of attributes. It might have made it possible to use bigger data sets with more attributes than those employed in this study to address other issues with the banking sector in the country. As a result, the next section gives some recommendations based on the research findings.

5.3 Recommendations

Even if the research investigation is primarily for academic purposes, it is going to have a significant impact on both the organization and other academics who are passionate about similar

topics. It emphasized the applicability of data mining technology to predict customers' lifetime value, which will aid in the organization's customer relationship management. The following suggestions are being provided to the organization and other interested researchers based on the research's findings.

- ✓ We encourage further research to connect a customer lifetime value predicting model with a CRM database management system based on the optimal model proposed in this study.
- ✓ Future researchers can use all 39 million datasets owned by CBE to develop a more beneficial customer lifetime value prediction model and increase the overall customer retention rate by knowing each customer's value and making strategies for retaining customers. However, the study was conducted with limited computing power, so additional research in the field is required using computers with high processing and memory capacity, capable of handling larger datasets, to improve the model.
- ✓ Since this study focuses on the application of data mining techniques, regression analysis for customer lifetime value prediction. So, the researchers recommend that other data mining approaches, such as classification and clustering, be investigated in the banking sector for predicting the lifetime value of customers.
- ✓ In addition to the deposit, the business should examine credit amount, debit amount, number of transactions, ATM card status, net monthly income and balance of customer which are key determinant variables to predict CLV based on the research findings. The company should identify customers based on their value determined by the modeling results, which predict how much value each customer will contribute to the organization.

Reference

[1] S. Chen1, Y. Huang, D.-L. Xu, and J. Wei, "A two stage machine learning approach for Modeling Customer Lifetime Value in the Chinese Airline Industry," 2020.

[2] H. Aeron, A. Kumar, and J. Moorthy, "Application of data mining techniques for customer lifetime value parameters : a review Harsha Aeron * Ashwani Kumar," no. March 2015, 2010, doi: 10.1504/IJBIS.2010.035744.

[3] K. Neha and M. Y. Reddy, "A study on applications of data mining," *Int. J. Sci. Technol. Res.*, vol. 9, no. 2, pp. 3385–3388, 2020.

[4] "Comercial Bank of Ethiopia about page - Commercial Bank Of Ethiopia." https://www.combanketh.et/en/about/ (accessed Dec. 19, 2022).

[5] E. Sharda, R., Delen, D. & Turban, *Business intelligence, analytics, and data science : a managerial perspective. Harlow, England: Pearson.* 2018.

[6] M. Völcker, C. Stenfelt, K. Skolan, and F. Teknikvetenskap, "Modelling Customer Lifetime Value in the Retail Banking Industry," 2021.

[7] Y. Kefetew, "Cluster analysis for customer segmentation in Commercial Bank of Ethiopia.," Addis Ababa University, 2021.

[8] B. Gebreegziabher and S. To, "Bank Customer Churn Prediction Model: the Case of Commercial Bank of Ethiopia," St.Mary University, 2022.

[9] W. DIBABA, "Application of Data Mining Techniques Management of Microfinance : the case of Wisdom Microfinance," Addis Ababa University, 2009.

[10] M. Haenlein, A. M. Kaplan, and A. J. Beeser, "A Model to Determine Customer Lifetime Value in a Retail Banking Context," *Eur. Manag. J.*, vol. 25, no. 3, pp. 221–234, 2007, doi: 10.1016/j.emj.2007.01.004.

[11] F. Games and P. Burelli, "Predicting Customer Lifetime Value in," pp. 1–28, 2019.

[12] NBE, "National Bank of Ethiopia Annual Bulletin," vol. 01, no. 1993, pp. 1–120, 2021.

[13] Cepheus Capital Research, “Ethiopia s Banking Sector,” vol. May 30, pp. 1–72, 2019.

[14] CBE, “Commercial Bank of Ethiopia,” *Commer. Bank Ethiop.*, vol. 23, no. 30, p. http://www.combanketh.et/, 2012, [Online]. Available: http://www.combanketh.et/.

[15] T. Sahoo, “Customer relationship management in Banks,” *Int. Res. J. Mod. Eng. Technol. Sci.*, vol. Volume:02/, no. March, pp. 99–130, 2020.

[16] L. P. Zhang, “Study on data mining for customer relationship management,” *Information, Comput. Appl. Eng. - Proc. Int. Conf. Inf. Technol. Comput. Appl. Eng. ITCAE 2014*, vol. 1, no. 6, pp. 921–924, 2015, doi: 10.1201/b18658-215.

[17] E. Nikumanesh and A. Albadvi, “Customer’s lifetime value using the RFM model in the banking industry: A case study,” *Int. J. Electron. Cust. Relatsh. Manag.*, vol. 8, no. 1–3, pp. 15–30, 2014, doi: 10.1504/IJECRM.2014.066876.

[18] D. R. Mani, J. Drew, A. Betz, and P. Datta, “Statistics and data mining techniques for lifetime value modeling,” pp. 94–103, 1999, doi: 10.1145/312129.312205.

[19] L. C. Yean and V. K. T. Khoo, “Customer relationship management: Lifecycle of predicting customer lifetime value,” *2nd Int. Conf. Comput. Res. Dev. ICCRD 2010*, pp. 88–92, 2010, doi: 10.1109/ICCRD.2010.24.

[20] M. R. Venkatakrishna, M. P. Mishra, M. Sneha, and P. Tiwari, “Customer Lifetime Value Prediction and Segmentation using Machine Learning,” *Int. J. Res. Eng. Sci. ISSN*, vol. 9, no. 8, pp. 36–48, 2021, [Online]. Available: www.ijres.org.

[21] S. Agarwal, *Data mining: Data mining concepts and techniques*, Third Edit. Morgan Kaufmann Publishers is an imprint of Elsevier., 2014.

[22] Y. Li, “Data Mining: Concepts, Background and Methods of Integrating Uncertainty in Data Mining,” *CCSC: SC Student E-Journal*, vol. 3, pp. 2–7, 2010.

[23] A. Azevedo and M. F. Santos, “KDD , SEMMA AND CRISP-DM : A PARALLEL OVERVIEW Ana Azevedo and M . F . Santos,” *IADIS Eur. Conf. Data Min.*, pp. 182–185, 2008, [Online]. Available: http://recipp.ipp.pt/handle/10400.22/136%0Ahttp://recipp.ipp.pt/bitstream/10400.22/136/3

/KDD-CRISP-SEMMA.pdf.

[24] T. Hendrickx, B. Cule, P. Meysman, S. Naulaerts, K. Laukens, and B. Goethals, “Mining association rules in graphs based on frequent cohesive itemsets,” *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 9078, no. 3, pp. 637–648, 2015, doi: 10.1007/978-3-319-18032-8_50.

[25] DataCamp, “【Pandas】Python for Data Science,” *Cheatsheet*, vol. 71, no. 10, pp. 1385–1389, 2001, [Online]. Available: https://www.datascience-pm.com/crisp-dm-2/.

[26] M. S. Brown, “What IT Needs To Know About The Data Mining Process,” *Forbes*, 2015. https://www.forbes.com/sites/metabrown/2015/07/29/what-it-needs-to-know-about-the-data-mining-process/?sh=2ce10d7e515f (accessed Apr. 11, 2023).

[27] “CRISP-DM is Still the Most Popular Framework for Executing Data Science Projects - Data Science Process Alliance.” https://www.datascience-pm.com/crisp-dm-still-most-popular/ (accessed Apr. 11, 2023).

[28] B. M. Ramageri, “Data mining techniques and applications,” *Indian J. Comput. Sci. Eng.*, vol. Vol. 1, p. No. 4 301-305.

[29] “Linear Regression in Machine learning - Javatpoint.” https://www.javatpoint.com/linear-regression-in-machine-learning (accessed Feb. 20, 2023).

[30] “Machine Learning Basics: Decision Tree Regression | by Gurucharan M K | Towards Data Science.” https://towardsdatascience.com/machine-learning-basics-decision-tree-regression-1d73ea003fda (accessed May 18, 2023).

[31] “Decision Tree Regression Explained with Implementation in Python | by The Click Reader | Medium.” https://medium.com/@theclickreader/decision-tree-regression-explained-with-implementation-in-python-1e6e48aa7a47#id_token=eyJhbGciOiJSUzI1NiIsImtpZCI6IjkxMWUzOWUyNzkyOGFlOWYxZTlkMWUyMTY0NmRlOTJkMTkzNTFiNDQiLCJ0eXAiOiJKV1QifQ.eyJpc3MiOiJodHRwczovL2FjY291bnRzLmdvb2dsZS5jb20iLCJhenAiOiIyMTYyOTYwMzU4MzQtazFrNnFlMDYwczJ0cDJhMmphbTRsamRjbXMwMHN0dGcuYXBwcy5nb29nbGV1c2VyY29udGVudC5jb20iLCJhdWQiOiIyMTYyOTYwMzU4MzQtazFrNnFlMDYwczJ0c

DJhMmphbTRsamRjbXMwMHN0dGcuYXBwcy5nb29nbGV1c2VyY29udGVudC5jb20i
LCJzdWIiOiIxMDE5NjE5NTAxOTE0MDg4NzA5NzIiLCJlbWFpbCI6InRydWFsZW1z
aXNheTg4QGdtYWlsLmNvbSIsImVtYWlsX3ZlcmlmaWVkIjp0cnVlLCJuYmYiOjE2O
TE0ODA4ODEsIm5hbWUiOiJUcnVhbGVtIFNpc2F5IiwicGljdHVyZSI6Imh0dHBzOi8v
bGgzLmdvb2dsZXVzZXJjb250ZW50LmNvbS9hL0FBY0hUdGNPZUFpRlVTVXlKYlh
uYkRCZUlmaWQxYVhOb2ZydVhrN0UzM2FqbkpJMj1zOTYtYyIsImdpdmVuX25hbW
UiOiJUcnVhbGVtIiwiZmFtaWx5X25hbWUiOiJTaXNheSIsImxvY2FsZSI6ImVuIiwiaW
F0IjoxNjkxNDgxMTgxLCJleHAiOjE2OTE0ODQ3ODEsImp0aSI6ImI0YTk3N2RhZjgz
ZDUzZjNiNGM3YmE0Y2RiODA3ZGU3ODk5ZDBkYmMifQ.rjv-S-
B6ei6JcdJq2GbFjtLNlu7NDTwh-
bAC_BjTayNPtSAZOQiz0SmSGcByZ0PnfU_hLNCTVdDxhJ82GBtZpxkfdyTPLptcbZ
WERpPuNGH61vonMtVkaLIPWe1t_-kUkaXoOmYne0nwm3-
vbF7EDIL7MIw4YR0slaish9_nmdVp8Guqh2hATyZcHaxm8bpPRbdOaNyplLZ1vk_482
stzXlo7fsPGfEDkknCgIcyppW84PXiUftiaq5wBAD1klSfN5QYtT7D7Hfs3fLDUP4DhQ
YoZPyb7r2-DH6OxF5QZWsf2HrYu7IvaEtJrkUF-p5mi7l0JvxbLGlAxrVvDS8lig
(accessed Aug. 08, 2023).

[32] "Understanding Random Forest. How the Algorithm Works and Why it Is… | by Tony Yiu | Towards Data Science." https://medium.com/p/58381e0602d2 (accessed Feb. 14, 2023).

[33] "Machine Learning Random Forest Algorithm - Javatpoint." https://www.javatpoint.com/machine-learning-random-forest-algorithm (accessed Feb. 23, 2023).

[34] A. Cutler, D. R. Cutler, and J. R. Stevens, "Ensemble Machine Learning," *Ensemble Mach. Learn.*, no. February 2014, 2012, doi: 10.1007/978-1-4419-9326-7.

[35] A. Jadon, A. Patil, and S. Jadon, "A Comprehensive Survey of Regression-Based Loss Functions for Time Series Forecasting," 2022, [Online]. Available: http://arxiv.org/abs/2211.02989.

[36] X. Dai, "Customer Lifetime Value Analysis Based on Machine Learning," *ACM Int. Conf. Proceeding Ser.*, pp. 13–17, 2022, doi: 10.1145/3546157.3546160.

[37] J. Friedrich, “Design science 97,” *AI Soc.*, vol. 10, no. 2, pp. 199–217, 1996, doi: 10.1007/BF01205282.

[38] “Recursive Feature Elimination: Working, Advantages & Examples.” https://www.analyticsvidhya.com/blog/2023/05/recursive-feature-elimination/ (accessed Oct. 01, 2023).

[39] “Chi-Square Test for Feature Selection in Machine Learning | by Sampath Kumar Gajawada | Towards Data Science.” https://towardsdatascience.com/chi-square-test-for-feature-selection-in-machine-learning-206b1f0b8223 (accessed Oct. 04, 2023).

[40] A. Marandon, “Using Machine Learning to Predict Customer Behaviour,” *Datasciencecentral.Com*, 2016.

[41] “Why Python Is Essential for Data Analysis and Data Science in 2021.” https://www.simplilearn.com/why-python-is-essential-for-data-analysis-article (accessed Feb. 12, 2023).

[42] A. Gholamy, V. Kreinovich, and O. Kosheleva, “Why 70/30 or 80/20 Relation Between Training and Testing Sets : A Pedagogical Explanation,” *Dep. Tech. Reports*, vol. 1209, pp. 1–6, 2018.

[43] M. A. Hall and G. Holmes, “Benchmarking Attribute Selection Techniques for Discrete Class Data Mining,” *IEEE Trans. Knowl. Data Eng.*, vol. 15, no. 6, pp. 1437–1447, 2003, doi: 10.1109/TKDE.2003.1245283.

[44] M. L. Sylvia and M. F. Terhaar, “Predictive Modeling,” *Clin. Anal. Data Manag. DNP*, pp. 10–13, 2023, doi: 10.1891/9780826163240.0024.

Appendix

A-1: CBE Dataset Information Prepared 19 Attribute for the Research

	CUSTOMER_CODE	AGE	GENDER	CUST_EDU	MARITAL_STATUS	DISTRICTNAME	ATM_CARD_STATUS	MOBILE_BANKING_STATUS
0	1045769235	23	MALE	Preparatory School	SINGLE	ASSELA	USER	USER
1	1045777668	43	FEMALE	Other	SINGLE	MERKATO	USER	USER
2	1045767863	31	MALE	First Degree	MARRIED	HAWASSA	NOT_USER	USER
3	1040893154	36	MALE	Elementary School	SINGLE	NEKEMTE	NOT_USER	NOT_USER
4	1040897050	25	FEMALE	First Degree	SINGLE	GONDER	USER	USER

A-2: CBE Dataset Information Prepared 19 Attributes for the Research cont.

INTERNET_BANKING_STATUS	OWNERSHIP	CATEGORY	DEBIT_AMOUNT	OPENING_DATE	NET_MONTHLY_IN	LAST_TXN_DATE	TOTAL_TXN	CREDIT_AMOUNT
NOT_USER	PRIVATE	SAVING	138750	8/25/2020	3000.0	12/8/2022	31	140050
NOT_USER	PRIVATE	SAVING	7618817	8/25/2020	5000.0	12/29/2022	314	7489486
NOT_USER	PRIVATE	SAVING	112000	8/24/2020	8017.0	12/30/2022	56	115499
NOT_USER	PRIVATE	SAVING	3600	10/5/2019	500.0	12/22/2022	11	3560
NOT_USER	PRIVATE	SAVING	13376	10/5/2019	600.0	12/31/2022	153	13081

A-3: CBE Dataset Information Prepared 19 Attributes for the Research cont.

LCY_CLOSING_BALANCE	CLV
1744	247.382431
1473	13405.770190
7312	201.683511
1218	4.931129
227	18.221074

B: Transforming Categorical Variables by Label Encoding

```
from sklearn.preprocessing import LabelEncoder
from sklearn import preprocessing

class Base_Feature_Engineering():

    def __init__(self):
        print("Feature Engineering object created")

        self.mapping_dict = {}

    #This method hepls to encode all the categorical variables with Labelencode

    def _Label_Encoding(self, data):
        category_cols = [col for col in data.columns if data[col].dtype == "object"]
        labelEncoder = preprocessing.LabelEncoder()

        for col in category_cols:
            data[col] = labelEncoder.fit_transform(data[col])
            le_name_mapping = dict(zip(labelEncoder.classes_, labelEncoder.transform(labelEncoder.classes_)))
            self.mapping_dict[col] = le_name_mapping

        return data
```

C: Categorical Variables after Transformed by Label Encoding

GENDER	CUST_EDU	MARITAL_STATUS	DISTRICTNAME	ATM_CARD_STATUS	MOBILE_BANKING_STATUS	INTERNET_BANKING_STATUS	OWNERSHIP	CATEGORY	[
1	8	4	3	1	1	0	2	1	
0	6	4	23	1	1	0	2	1	
1	3	1	14	0	1	0	2	1	
1	2	4	25	0	0	0	2	1	
0	3	4	12	1	1	0	2	1	

D: Numerical Attributes Count Summary

	Negative values count	Positive values count	Zero count	Unique count	Negative Infinity count	Positive Infinity count	Missing Percentage	Count of outliers
CUSTOMER_CODE	0.0	100101.0	0.0	95430.0	0.0	0.0	0.0	18539.0
AGE	0.0	100101.0	0.0	96.0	0.0	0.0	0.0	3972.0
DEBIT_AMOUNT	0.0	99956.0	145.0	46778.0	0.0	0.0	0.0	13623.0
NET_MONTHLY_IN	0.0	98098.0	2003.0	4618.0	0.0	0.0	0.0	7530.0
TOTAL_TXN	0.0	100101.0	0.0	1027.0	0.0	0.0	0.0	10867.0
CREDIT_AMOUNT	0.0	100074.0	27.0	49681.0	0.0	0.0	0.0	13695.0
LCY_CLOSING_BALANCE	0.0	99357.0	744.0	25214.0	0.0	0.0	0.0	16500.0
CLV	0.0	100075.0	26.0	96840.0	0.0	0.0	0.0	14186.0

E: Schema of Data Frame with Attribute Names, Types, Missing Values and Sample Observations

```
=============================================================================================================
FEATURE NAME     DATA TYPE        # OF MISSING VALUES  SAMPLES
CUSTOMER_CODE      int64               0                  1045769235,1045777668,1045767863,1040893154,1040897050,
CATEGORY           object              0                  SAVING,SAVING,SAVING,SAVING,SAVING,
LCY_CLOSING_BALANCE int64               0                  1744,1473,7312,1218,227,
CREDIT_AMOUNT      int64               0                  140050,7489486,115499,3560,13081,
TOTAL_TXN          int64               0                  31,314,56,11,153,
LAST_TXN_DATE      object              0                  12/8/2022,12/29/2022,12/30/2022,12/22/2022,12/31/2022,
NET_MONTHLY_IN     float64             0                  3000.0,5000.0,8017.0,500.0,600.0,
OPENING_DATE       object              0                  8/25/2020,8/25/2020,8/24/2020,10/5/2019,10/5/2019,
DEBIT_AMOUNT       int64               0                  138750,7618817,112000,3600,13376,
OWNERSHIP          object              0                  PRIVATE,PRIVATE,PRIVATE,PRIVATE,PRIVATE,
AGE                int64               0                  23,43,31,36,25,
INTERNET_BANKING_STATUS object             0                  NOT_USER,NOT_USER,NOT_USER,NOT_USER,NOT_USER,
MOBILE_BANKING_STATUS object            0                  USER,USER,USER,NOT_USER,USER,
ATM_CARD_STATUS    object              0                  USER,USER,NOT_USER,NOT_USER,USER,
DISTRICTNAME       object              0                  ASSELA,MERKATO,HAWASSA,NEKEMTE,GONDER,
MARITAL_STATUS     object              0                  SINGLE,SINGLE,MARRIED,SINGLE,SINGLE,
CUST_EDU           object              0                  Preparatory School,Other,First Degree,Elementary School,First D
egree,
GENDER             object              0                  MALE,FEMALE,MALE,MALE,FEMALE,
CLV                float64             0                  247.3824312,13405.77019,201.6835106,4.931129477,18.22107438,
=============================================================================================================
```

F: Removing Outliers

```
    def remove_outliers(self,data):

        """
        This method helps
        to remove the outliers
        from the target
        variable, hence it
        removes the influencial
        values
        """

        q1 =df['CLV'].quantile(.25)
        q3 = df['CLV'].quantile(.75)
        iqr = q3-q1
        df_out = df[~((df['CLV'] < \
        (q1 - 1.5 *iqr))  |  (df['CLV'] > \
                              (q3+ 1.5 * iqr)))]
        return df_out
```

G: Model Building using Linear Regression

```
import statsmodels.api as sm

class Data_Modelling():


    def __init__(self,n_estimators=100,random_state=42,max_depth=10):

        print("Data Modelling object created")


    def OLS_Summary(self,data):
        model2 =sm.OLS(y_train,x_train).fit()
        return model2.summary()


    def Linear_Regression_Model(self,df):
        regressor = LinearRegression()
        reg=regressor.fit(x_train,y_train)
        LR_pred=regressor.predict(x_test)
        LR_RMSE = np.sqrt(metrics.mean_squared_error(y_test,LR_pred))
        LR_r2_score = r2_score(y_test,LR_pred)
        return LR_RMSE,LR_r2_score
```

H: Decision Tree Regression Model

```
def Decision_Tree_Model(self,df):
    regressor = DecisionTreeRegressor(random_state=29)
    reg=regressor.fit(x_train,y_train)
    DT_pred=regressor.predict(x_test)
    DT_RMSE = np.sqrt(metrics.mean_squared_error(y_test,DT_pred))
    DT_r2_score = r2_score(y_test,DT_pred)
    return DT_RMSE,DT_r2_score
```

I: Random Forest Regression Model

```
def Random_Forest_Model(self,df):
    regressor = RandomForestRegressor(n_estimators=100,random_state=29,max_depth=12)
    reg=regressor.fit(x_train,y_train)
    RF_pred=regressor.predict(x_test)
    RF_RMSE = np.sqrt(metrics.mean_squared_error(y_test,RF_pred))
    RF_r2_score = r2_score(y_test,RF_pred)
    return RF_RMSE,RF_r2_score
```