CHAPTER I

INTRODUCTION

1.1. Motivation

To remain competitive, businesses must adapt to market demands and prioritize customer needs. As companies grow and gain popularity, predicting market value has become essential for sustaining success. Customer Relationship Management (CRM) has significantly improved profitability in recent years through strategies that enhance customer retention and purchases.[1],[2] A key CRM technique is Customer Lifetime Analysis (CLA), which helps businesses understand customer preferences and make informed decisions using data. Ensemble machine learning combines multiple models to deliver more accurate and stable predictions.

This research focuses on CLA in the retail sector, where it identifies customers likely to make repeat purchases and aids in developing targeted marketing campaigns to boost loyalty. CLA also helps optimize pricing strategies by identifying customers sensitive to price changes.[3] In Indonesia's motorcycle industry, CLA enables businesses to assess customer value, supporting decisions on acquisition, retention, and loyalty programs.

Market value prediction and CLA are essential tools for identifying profitable customer segments and tailoring marketing and sales strategies. Accurate predictions improve inventory management,

pricing, and marketing while uncovering growth opportunities. These insights enable businesses to develop strategies that capitalize on market dynamics and foster long-term success.

In Customer Relationship Management (CRM), Customer Lifetime Value (CLV) is a critical metric for evaluating marketing decisions.[4] Combining CLV analysis with market value prediction using ensemble machine learning provides businesses with a competitive edge. Accurate CLV predictions enable businesses to optimize operations and marketing strategies, leading to increased revenue and profitability. Since data has varying strengths and limitations, developing accurate and reliable CLV predictions is essential.

This study integrates CLV analysis with ensemble machine learning models to improve prediction accuracy. Ensemble learning captures diverse patterns and relationships in data, resulting in more precise CLV predictions. The models utilized in this research focus on XGBoost and Random Forest. Table 1.1 outlines the advantages and disadvantages of these machine learning methods.

Table 1 Machine Learning Models Advantages and Disadvantages

Machine Learning Methods	Advantages	Disadvantages

Random Forest Regression	Handles nonlinearity, reduces overfitting, provides feature importance, robust to missing data	Computationally intensive, less interpretable, may struggle with imbalanced data
XGBoost	High accuracy, manages missing data, scalable, customizable, reduces overfitting	Complex to tune, computationally expensive, less interpretable, can overfit if not finely tuned

The dataset used in this study originates from invoice records of a motorcycle business in Sumatra, Lampung. This data is categorized as non-contractual, as it involves ad-hoc transactions without prior lease agreements.

CLV measurement focuses on two primary contexts: **non-contractual** and **contractual**. While both are important for Customer Lifetime Analysis (CLA), the choice of context depends on the nature of the available data. In this case, the focus is on the non-contractual context, where customer behaviour is not tied to formal agreements.[5],[6]

In a non-contractual context, firms do not observe customer defection, and the relationship between purchase behaviour and

customer lifetime is uncertain.[7],[8] Conversely, in a contractual context, customer defections are observable, and a longer customer lifetime directly indicates a higher CLV.[8]

Given the nature of the dataset, it aligns with the non-contractual context. This assumption is based on the ad-hoc relationships observed in the daily business activities reflected in the data.

Customer Lifetime Analysis (CLA) is essential for aligning customer interests with business continuity and making informed marketing decisions. Calculating Customer Lifetime Value (CLV) involves three key steps:[9]

1. Determine average order.
2. Calculating average number of transactions per period.
3. Measure customer retention.

While CLV analysis provides valuable insights into market and customer behavior, traditional methods often fall short in accurately predicting trends. To address this, machine learning offers data-driven predictive capabilities. This research focuses on applying Ensemble Machine Learning Models to transactional data to develop accurate CLV prediction models. These models enable actionable analysis to enhance customer engagement and improve Customer Relationship Management (CRM) while increasing CLV.

1.2. Problem Formulation

Predicting Customer Lifetime Value (CLV) is crucial for enhancing customer management and long-term profitability. Accurate predictions help businesses identify valuable customers, refine marketing strategies, and allocate resources efficiently. However, the motorcycle industry faces unique challenges in building reliable CLV models:

1. Transactional data, like tax invoices, often lacks detailed customer behaviour or demographic information, limiting insight into factors influencing CLV.
2. Customer purchasing patterns are complex and nonlinear, making traditional models like Linear Regression inadequate.
3. Choosing the right machine learning model requires balancing accuracy, efficiency, and interpretability.
4. Identifying key features (e.g., transaction frequency, total spending) is vital but difficult with advanced models that lack transparency.
5. Evaluating model performance on transactional data is essential, but ensuring generalization to unseen data or other industries remains a challenge.

1.3. Scopes of Problem

This study examines predicting Customer Lifetime Value (CLV) in the motorcycle industry using Random Forest Regression and XGBoost. It evaluates the accuracy of these models in providing

actionable insights for customer management but faces the following limitations:

1. The dataset includes two years of transactional tax invoice data, which may not fully capture broader customer behavior or industry nuances, limiting the quality of insights.
2. The dataset excludes demographic, behavioural, and external factors, limiting the model's scope, but it will be cleansed and normalized to minimize deviations.
3. The dataset lacks temporal trends, such as changing preferences, market dynamics, or economic conditions, which may impact prediction accuracy.
4. The analysis is limited to two machine learning models: Random Forest Regression and XGBoost.

1.4. Research Purpose

The growing use of Customer Lifetime Value (CLV) in recent years highlights its importance. This research aims to evaluate and compare the performance of various machine learning models in predicting CLV using tax invoice datasets. By leveraging these techniques, the study seeks to identify the most effective model for accurately predicting CLV, a vital metric for strategic decision-making in customer relationship management.

1.5. Research Methodology

The research methodology consists of four key steps:

1. **Literature Review**: The study begins with a review of existing approaches to Customer Lifetime Analysis (CLV), identifying their limitations and areas for improvement.
2. **Data Analysis and Preparation**: Researchers analyse the dataset, which consists of invoice logs for motorcycle sales, adopting a noncontractual context due to uncertain relationships between purchase behaviour and customer lifetime. Relevant attributes are selected, and the dataset is cleansed and normalized to minimize errors.
3. **Model Development**: Ensemble learning models, specifically tailored to the data and context, are developed to meet the requirements of CLV prediction.
4. **Training, Prediction, and Evaluation**: The developed models are trained, used for prediction, and evaluated using appropriate metrics. Performance metrics will help assess the accuracy and reliability of the models, enabling a clear analysis of CLV and identification of key influencing factors.

This structured approach ensures a comprehensive analysis of CLV, providing clear insights into model performance and the factors affecting predictions.