DETAILED MPR DOCUMENT

MSC_DISSERTATION_ _V1.9

COIS71052-2023-SPG2

CUSTOMER LIFETIME VALUE PREDICTION IN SUBSCRIPTION SERVICES

AUGUST 04, 2024

ABSTRACT

Humans prioritize wants primarily based on the satisfaction or values they derive from them. Organizations are in business to provide these wants, either in form of goods or services, and make profits. It is therefore imperative for organizations to monitor their customers' behavioural dynamics. One critical metric that offers insights on the value a customer contributes to a company is the customer lifetime value (CLTV). This study explored the complex dynamics of Customer Lifetime Value (CLV) prediction in subscription-based businesses, aiming to research into customer lifetime value in subscription-based businesses, to develop predictive models that accurately forecasts customer lifetime value, to implement the predictive models to analyze the business data, including the RFM (Recency, Frequency, and Monetary) model, Markov chain, and Pareto/NBD model in predicting Customer Lifetime Value (CLV), to test these models with new data for validation.. Previous studies have often ignored the probabilistic nature of subscription-based businesses, churn and engagement attributes, customer retention which is one of the goals of customer relationship management. The research journey covered a comprehensive review of literature, methodological framework development, data analysis, and model development, culminating in a rich distillation of insights into CLV dynamics and their implications for subscription services. Model development emerges as the core of this research, with a focus on three key methodologies: RFM analysis, Markov Chain modeling, and Pareto/NBD modeling. These techniques offered unique insights into customer segmentation, segment transition prediction, and CLV estimation, empowering businesses to tailor their marketing strategies and retention efforts to maximize customer lifetime value. The dissertation concludes with a set of actionable recommendations for subscription-based businesses seeking to enhance their CLV prediction capabilities and optimize customer engagement.

CHAPTER ONE: INTRODUCTION

1.0 Introduction

Subscription based businesses are businesses running on a model that requires recurring payments at stipulated intervals, namely, daily, weekly, monthly, annual, or seasonal for delivery of products or services to customers. These business models have achieved significant progress across various industries. Many organizations have adjusted from traditional business model of one-time payment to recurring payments strategies which characterizes subscription-based business model on account of globalization and evolution of technologies (Lindstrom et. al., 2023). Businesses are inclined to this model as it offers them the benefits of knowing accurately the number of their customers or clients that are active or recently churned (McCormick, 2016), and potentially the revenue that accrues from customers' subscriptions. However, beyond these benefits, a more crucial imperative for businesses adopting this model is customer relationship management (Subbly Blog, 2022). Subscription-based services also offer advantages to the customers or subscribers. Such benefits range from flexibility where the subscriber can churn, the subscriber is not obliged to make upfront payment, feeling of exclusivity, from a social standpoint.

However, these subscription-based businesses are not without challenges, despite the benefits. Challenges of subscription-based models include scalability in logistics without necessarily inducing higher cost, value sustainability, optimal pricing strategies, and process automation. More than anything else, subscribers want values for their spending, and businesses must be aware of this disposition and factor that in in their pricing strategies and structure, as these are important factors customers consider in their decision making. But then this may create a complex situation as the values offered by the vendor must also be sustainable. Seamless and automated customer relationship management processes like sending appreciation or promo-awareness text messages, emails to customers that reached certain benchmarks, choosing a robust platform online payment, outsourcing logistics operations where necessary could prove impactful.

1.1 Research Background

Humans prioritize wants primarily based on the satisfaction or values they derive from them. This is what economists refer to as utility. Organizations are in business to provide these wants, either in form of goods or services, and make profits. It is therefore imperative for organizations to monitor their customers' behavioural dynamics. One critical metric that offers insights on the value a customer contributes to a company is the customer lifetime value (CLTV). In most literatures, it is often measured in terms of average order value, purchase frequency, and retention length. CLTV is basically a metric that identifies customers with the greatest potential net value over the period the customers remain active (Ramachandran, 2006).

Understanding and predicting Customer Lifetime Value (CLV) is critical for subscription-based businesses to retain customers and ultimately maximize revenue. Chang, et. al. (2012) stated that CLV is a crucial factor that managers consider when they want to evaluate the value offered by each customer and assess a firm's performance because there is a direct linkage between financial returns and marketing efforts. CLV estimates the potential value a customer will offer in monetary terms for as long as he remains loyal to a business. Accurately predicting this CLV for subscription-based businesses is crucial for maximum profitability and sustainability. With evolutions of technologies, CLV can be predicted more accurately and efficiently leveraging sophisticated models and methodologies. For example, Singh (2023), in his study, had applied recency, frequency, and monetary (RFM) model to predict CLV. He established the effectiveness of leveraging predictive modeling of CLV to provide competitive edge for businesses. This study attempts to explore and critically evaluate the effectiveness of the RFM model and compare it with the BG/NBD and Pareto/NBD models.

While there has been an increasing interest in models that can accurately predict CLV, finding a one-size-fits-all model remains elusive. Previous studies have often ignored the probabilistic nature of subscription-based businesses, churn and engagement attributes, customer retention which is one of the goals of customer relationship management. More often than not, they base the calculation of CLV on average order value, frequency of purchase, and the life span of the customer, leaving out churn and customer engagement which are directly related to customer relationship management and ultimately the value a business can derive from them. Addressing these gaps is crucial for improving predictive accuracy and scalability; thus, this study not only aims to develop and implement models that can predict the future values business might obtain from their subscribers but to uncover the intrinsic dynamics that characterize subscribers' behaviour and engagement pattern. This is crucial because most studies have always based their methodologies on predicting the CLV with respect to whether the subscribers are active or inactive but this study goes beyond this to uncover the transition dynamics of customers which will ultimately inform businesses on the strategies to deploy in order to foreclose imminent churn. The researcher believes this is the best approach to create or increase lifetime values of subscribers in businesses.

This research will obtain historical customer data which shall, among other things, include transactional records, subscription plans, demographics, and engagement metrics, from relevant subscription-based services. These data shall be cleaned, transformed, modeled, and preprocessed following standardized approach for handling missing values, outliers, and inconsistencies, to ensure data quality and integrity is achieved.

1.2 Research Questions and Hypotheses

1.2.1 Research Question

- What predictive model will best capture customer demographics, transactions pattern, customer engagements, and accurately forecasts customer lifetime value (CLV) in subscription-based businesses?

1.2.2 Hypotheses:

- The most effective model will capture customer demographics, transactions pattern, customer engagements, and accurately forecasts customer lifetime value (CLV) in subscription-based businesses.

1.3 Research Aim

The aim of this study is to research into customer lifetime value in subscription-based businesses, to develop predictive models that accurately forecasts customer lifetime value, to implement the predictive models to analyze the business data, to critically evaluate and compare the effectiveness of different predictive models, including the RFM (Recency, Frequency, and Monetary) model, Beta Geometric/Negative Binomial Distribution model, and Pareto/NBD model in predicting Customer Lifetime Value (CLV), to test these models with new data for validation.

1.4 Research Objectives:

1. To investigate existing literature and methodologies on customer lifetime value (CLV) in subscription-based businesses.
2. To develop and refine predictive models that accurately and reliably forecasts customer lifetime value (CLV) for subscription-based businesses.
3. To identify key insights derived from the analysis of the predictive models for customer lifetime value (CLV) in subscription-based businesses.
4. To implement and analyze the accuracy and effectiveness of the RFM (Recency, Frequency, and Monetary) model, Beta Geometric/Negative Binomial Distribution model, and Pareto/NBD model, in predicting customer lifetime value (CLV) in subscription-based businesses.
5. To critically compare and evaluate the strengths, weaknesses, and limitations of each predictive model, considering factors such as computational complexity, data requirements, and interpretability.

6. To validate the predictive models using new data to assess their generalizability and robustness in real-world scenarios.

1.5 Significance and Justification of the Study

The contributions from this research go beyond normal academic discourse, infiltrating the landscape of subscription-based services with benefits transcending theoretical frameworks, engendering transformations in predictive analytics and strategic decision-making. It represents a critical point in the emergence of innovative subscription-based business models with far-reaching implications for academia, industry, and business stakeholders. By seamlessly integrating time series predictive analysis techniques with probabilistic models. It redefines customer lifetime value (CLV) prediction landscape with critical changes that decision-making processes across a broad spectrum of domains.

In the academia, this study represents a critical milestone in marketing analytics with regards to the innovative methodologies that expands the theoretical depth and concepts of subscription based CLV prediction. While these approaches are not exhaustive, the study further creates avenues to explore and refine further implications for subscription based predictive analytics and CLV estimation.

The implications of this research are incontrovertible from an industry standpoint. Subscription based businesses will benefit greatly with respect to insights generated from the advanced predictive analytics techniques. This will help companies gain specific understanding of customer behaviours and preferences, and consequently customize their acquisition, retention, and pricing strategies accordingly. This study will also empower industry practitioners with the right tools and methodologies to explore the intricate landscape of subscription-based business models for competitiveness and sustainable growth. Business executives, marketers, analysts, and other stakeholders will benefit immensely from the transformative findings of this study in terms of making better data-driven decisions, informed resource allocation, investment preferences and prioritization, and customer engagement efforts, with incontrovertible accuracy.

1.6 Scope and Limitation of the Study

The scope of this research covers a detailed exploration of predictive analytics techniques and probabilistic models for CLV prediction within business models that are subscription-based. It focuses on the accuracy and reliability of CLV estimation leveraging time series analysis and probabilistic models thereby empowering businesses to make informed decisions with respect to customer acquisition and retention, and pricing strategies. Crucial to this study is the integration of theoretical frameworks and empirical evidence to develop sophisticated predictive models that will effectively handle nuanced dynamics of subscription services. Moreover, the study adopts a comparative approach, analyzing the accuracy and effectiveness diverse predictive analytics techniques in CLV estimation and pinpointing best practices for implementation.

While this research aims to provide detailed insights into CLV prediction within subscription services, it is important to recognize the inherent limitations that characterises the research landscape. One of such limitations has to with data availability and the quality of data. The accuracy and effectiveness of predictive models relies largely on the availability of accurate data sets that captures the relevant features of interest. Such data are proprietary customer data but the challenges with them are concerns about data privacy and security. Moreover, the dynamic nature of subscription-based services create intricacies that may confound predictive modeling efforts. Changing consumer behaviours and market forces are factors that necessitate a non-rigid and adaptive approach to CLV prediction. While this study considers these factors, uncertainties may still arise to forfeit the predictive accuracy of the models so developed. Furthermore, the research is subject to methodological limitations characteristic of predictive analytics and probabilistic modeling. Assumptions used for these predictive models may not always conform to real-world conditions, leading to errors between predicted outcomes and observed attributes. Also, the intricacies of probabilistic frameworks may pose challenges in model interpretation and validation, necessitating meticulous calibration and sensitivity analysis to ensure robustness and reliability.

1.7 Overview of the Dissertation Structure

The research is organized into six chapters, each addressing specific aspects of the research. Chapter One gives an overview of the research topic, problem statement, objectives, significance, scope, and structure of the dissertation. Chapter Two reviews theoretical frameworks, methodological approaches, and empirical studies bordering on CLV prediction in subscription-based business landscape, time series analysis, and probabilistic models. Chapter Three describes the research methodology which includes data collection strategies, time series frameworks, model development and evaluation criteria, and data analysis methods. Chapter Four describe the dataset(s) used, data gathering processes, cleaning and preprocessing efforts, handling of missing values and outliers, and overview of the final dataset for analysis. Chapter Five details the analytical approach, descriptive analysis of the dataset, application of predictive analytics techniques, development of CLV prediction models, model selection, and evaluation criteria. Chapter Six gives summary on the key findings, discusses implications for theory and practice, and provides useful recommendations for future research and practical applications in the field.

CHAPTER TWO: LITERATURE REVIEW

2.1 Introduction

The accurate prediction of customer lifetime value (CLV) in subscription-based businesses has emerged as a interesting area of research. Subscription models, characterised by recurring revenue streams and subscriber retention focus, have proliferated across various sectors, driven by the need for predictable income and deeper subscriber engagement. The successful prediction of CLV not only influences strategic decision-making but also impacts marketing strategies, resource allocation, and overall business sustainability. This literature review delves into the architecture, development, implementation, and evaluation of predictive models for CLV, with a particular focus on Recency, Frequency, and Monetary (RFM) analysis, the Beta-Geometric/NBD (BG/NBD) model, the Pareto/NBD model, and the Markov Chain model.

The theoretical framework underpinning this review is rooted in economic value theory, which provides a foundation for understanding how customer interactions translate into economic value and how predictive models can quantify this value. Economic value theory highlights the importance of considering future potential earnings when evaluating customer relationships, thus emphasizing the need for robust and accurate CLV predictions. For the conceptual framework, the review incorporates both subscriber demographics and socio-economic constructs into the predictive modeling framework. Subscriber demographics, such as age, gender, and location, play a crucial role in shaping customer behaviour and, consequently, their lifetime value. Socio-economic factors, including income level, education, and occupation, further influence purchasing decisions and engagement patterns. Integrating these factors into predictive models allows for a more streamlined understanding of subscriber behaviour and enhances the accuracy of CLV predictions.

Despite advancements in predictive modeling, significant research gaps remain. Many existing studies have predominantly focused on traditional models such as RFM analysis, which, while useful, often fall short in capturing the probabilistic nature of subscriber behaviour and state transitions. The RFM model, based on historical transaction data, provides valuable insights into subscriber segments but lacks the dynamic capability to account for changing subscriber states over time. Similarly, while the BG/NBD and Pareto/NBD models offer more sophisticated approaches by incorporating the probability of subscriber dropout and transaction frequency, they still have limitations in accounting for the evolving nature of subscriber relationships. This is where the Markov Chain comes handy to better address the issues of state transitions of subscribers over time. This model provides a framework for understanding how subscribers move between different states, for example, first-time subscribers to loyal subscribers to elite subscribers, and how these transitions impact their overall lifetime values. Despite its advantages, the application of Markov Chains in CLV prediction is still underexplored in the literature, presenting a notable research gap.

This review will critically examine the strengths and limitations of the RFM, BG/NBD, Pareto/NBD, and Markov Chain models, highlighting their contributions to CLV prediction and identifying areas for further study. By addressing these existing gaps and integrating advanced probabilistic methods, this review aims to put forward the understanding of CLV in subscription-based businesses and provide insights into improving predictive accuracy. Through a comprehensive examination of these models, the review will contribute to a more refined and effective approach to predicting and managing customer lifetime value in the evolving subscription business landscape.

2.2 Customer Lifetime Value and Subscription-based Businesses

Jasek et. al. (2018) had established that customer lifetime value represents a critical strategy in modern marketing, providing useful insights into the long-term profitability of customers acquisition, retention, and segmentation. There is a growing need to understand and leverage CLV with respect to market dynamics if businesses must thrive. CLV represents the total value or worth a subscriber offers a business over the lifespan of their relationship. As established by Sridhar and Corbey (n.d.) and Singh (2023), it is often measured as the difference between all revenues, present and future, made from the subscriber and the cost of acquisition, servicing, and retention of that subscriber. However, the conceptualization of CLV goes beyond mere financial quantification. It also includes establishing a healthy relationship with the subscriber. Over the years, CLV has evolved as a decisive metric that goes beyond traditional transactional analyses to embrace wider strategic requirements. Its evolution showcases the paradigmatic move from transactional to relational marketing (Chen and Fan, 2013), emphasizing the transition towards customer-centric business models. Pivotal to the significance of CLV is its role in dictating marketing strategies and ensuring long-term customer relationships (Wang et. al., 2019). By quantifying the long-term value of individual subscribers, businesses can target marketing efforts to maximize profitability and boost customer satisfaction. Moreover, CLV serves as a guide for resource allocation, channeling investment decisions towards initiatives that yield the highest returns over the customer lifecycle. By tracking dynamics in CLV metrics over time, businesses can measure the impact of various improvements on customer loyalty (Chamberlain et. al., 2017), satisfaction, and lifetime value. This iterative feedback loop ensures continuous improvement of marketing strategies, fostering a culture of data-driven decision-making within organizations.

Subscription-based business models have helped to birth a paradigm shift in consumer preferences and behaviours, creating a mindset of access over ownership and experiential consumption over material possession (Lindstrom et. al., 2023). This shift is exemplified by the rise of subscription-based streaming platforms like Netflix and Spotify (Vishkaei and Giovanni, 2023) which have revolutionized how individuals consume and interact with media content. Subscription services have engendered a reawakening in customer engagement efforts, leveraging recurring revenue streams to establish deeper, more lasting relationships with customer base. Through personalized recommendations, loyalty programs, and community-building initiatives, subscription businesses develop a sense of belonging and exclusivity among subscribers, thereby enhancing customer retention and lifetime value. However, the transformative potential of subscription-based business models transcends consumer behaviour to embrace wider market dynamics. By facilitating recurring revenue streams and predictable cash flows, subscription services draw resilience and agility upon businesses, eliminating the volatility inherent in traditional transactional models. This resilience is particularly critical in times of economic uncertainty, where subscription businesses are well positioned to cushion economic downturn.

2.3 Review Of CLV Prediction Models

CLV prediction models are crucial tools in the marketing and sales analytics landscape. Businesses leverage them for the optimization of their customer acquisition and retention strategies. CLV, from inception, had relied on simplistic heuristic models like the RFM (Recency, Frequency, Monetary) framework (Hiziroglu et. al., 2018), which offered basic insights into customer behaviour but lacked predictive accuracy and scalability. Over the years, the ecosystem of CLV prediction has undergone a significant transformation, fueled by advancements in data analytics, machine learning, and computational algorithms. Contemporary models were robust predictive frameworks with the ability to predict future customer behaviour with significant accuracy. However, notwithstanding the feats attained in CLV prediction, drawbacks ranging from data sparsity and heterogeneity to model interpretability and scalability abound. Traditional models, while intuitive and are easy to interpret, often fail to detect the intricate interplay of variables that influence customer lifetime value. Conversely, sophisticated machine learning models, while good at handling high-dimensional data, may suffer from overfitting and generalization issues, forfeiting their predictive performance in real-world scenarios.

Robust methodologies that incorporate RFM (Recency, Frequency, Monetary), Pareto/NBD, and Markov chain models represent a significant transformation in the field of Customer Lifetime Value (CLV) prediction, providing customized methodologies geared towards the unique characteristics of customer behaviour and purchase patterns. These methodologies, rooted in statistical modeling and probability theory, enable businesses to gain deeper insights into customer dynamics, forecast future behaviour, and optimize marketing strategies to maximize CLV. RFM analysis is used to segment subscribers based on their recent purchase behaviour, frequency of purchases, and monetary value derived from the transactional engagements of the subscribers (Khajvand, et. al., 2011). It is established on the notion that subscribers purchasing patterns can be effectively summarized by these three key metrics: recency, frequency, and monetary value. By segmenting subscribers based on these features, businesses can identify high-value segments of their subscriber base, finetune their target marketing strategies, and optimize resource allocation to maximize CLV. Recency refers to how recently a customer made a purchase. Subscribers who have made recent purchases are considered to have a higher likelihood to respond to transactional engagements and marketing efforts, an indication of higher potential for future purchases. Frequency measures how often a customer makes purchases within a specific timeframe. Subscribers with a higher purchase frequency are likely to be more loyal and valuable to the brand. Monetary value represents the amount of money spent by a subscriber on purchases; thus, subscribers with higher monetary value are most profitable to the brand, making them valuable targets for retention and upselling. RFM analysis classifies subscribers into distinct groups based on their RFM scores calculated as quantiles or percentiles of each RFM metric. For example, subscribers with the highest recency scores, indicating recent purchases, may be classified as "active" or "high-value" or “loyal” subscribers, while those with the lowest recency scores, representing no recent purchases, may be classified as "inactive" or "at-risk" or “disloyal” subscribers. In the same vein, customers with the highest frequency and monetary value scores may be classified as "top-spenders" or “high-value” subscribers. Business can then target their marketing strategies according to each class’ unique characteristics and preferences. For example, "active" subscribers may receive personalized offers or rewards to encourage repeat purchases, while "at-risk" customers may receive targeted reactivation campaigns to win them back. By leveraging RFM segmentation strategies, businesses can maximize the effectiveness of their marketing efforts, improve customer retention, and ultimately increase CLV.

The Pareto/NBD model is a probabilistic framework used to predict future purchase frequency and subscriber churn based on historical transaction data. The nomenclature is derived from two components, namely the Pareto distribution, which models subscriber inter-purchase times, and the Negative Binomial distribution, which models subscribers churn behaviour. The assumption in the application of this model is that subscribers’ purchasing behaviour follows a gamma-gamma distribution for monetary value and a Poisson or negative binomial distribution for transaction count. The probabilistic nature of Pareto/NBD model enables businesses to estimate the likelihood of future subscriber behaviour based on observed patterns, trends and historical data by providing valuable insights into customer lifetime value, retention rates, and future purchase probabilities. In reality, subscribers are never the same and will not behave the same way, creating a condition of heterogeneity and variability in purchasing behaviour. This scenario makes traditional models incapable of accurately predicting subscribers’ behaviour, making the Pareto/NBD model an indispensable tool in this regard. By simulating different scenarios and strategies, businesses can pinpoint the most effective methodologies for maximizing CLV and profitability. For example, businesses can accurately evaluate the impact of targeted promotions, loyalty programs, and subscriber engagement initiatives on subscriber retention and lifetime value.

Markov chain model is another probabilistic framework but finds its application in modeling subscribers transitions between different states, such as active, inactive, and churned. This mode, as the name implies, has its essence from the Markov property, a principle which states that future states depend only on the current state and are independent of past states. It is effective in capturing the stochastic nature of subscriber behaviour and the dynamics of subscriber-state transitions, thereby helping businesses to predict future CLV and optimize marketing strategies to maximize subscriber retention and lifetime value. In a Markov chain model, subscribers are represented as nodes in a state-transition graph, with edges representing the probabilities of transitioning between states. For example, a subscriber may transition from an active state to an inactive state with a certain probability, or from an inactive state to a churned state with another probability. With insights from historical transaction data, Markov chain model can aid businesses to estimate the transition probabilities and construct a Markov chain model to predict future subscriber behaviour. Even more interesting is the ability of Markov chain model to capture the temporal dynamics of subscriber behaviour and state transitions associated with market conditions, unlike static models that assume constant transition probabilities over time. This enables businesses to fashion their marketing strategies in response to evolving subscriber behaviour and market dynamics, thereby maximizing CLV and profitability.

2.4 Theoretical Framework

Crucial to this review is the theoretical foundation of subscriber behaviour, which underpin the conceptualization of CLV and its predictive models. Leveraging insights from psychological, sociological, and behavioural theories, the theoretical framework elucidates the drivers, motivations, and decision-making processes behind subscribers' interactions with businesses. In relationship marketing, long-term subscriber relationships are pivotal to business success. In this review, principle of economic value theory (EVT) and the strategic imperatives of establishing long-lasting and loyal subscriber base will be highlighted. By nurturing trust, loyalty, and reciprocity, businesses can grow a loyal subscriber base, thereby maximizing CLV and driving sustainable growth. Thus, through critical investigation, this review seeks to bridge the gap between theory and practice, fostering theoretical frameworks that guide the development and finetuning of CLV prediction models. By integrating insights from diverse disciplines, businesses can gain a holistic understanding of customer lifetime value and its strategic implications for subscription services.

Moreover, economic theories offer valuable insights into the dynamics of CLV prediction, highlighting the interactions between supply and demand, pricing strategies, and market dynamics. From traditional theories of consumer behaviour, such as the utility theory and price elasticity, to contemporary paradigms of behavioural economics and game theory, a rich reservoir of economic theories fortifies the understanding of CLV prediction and its implications for subscription-based business models. Furthermore, the discourse transcends theoretical abstraction to cover practical applications, demonstrating how theoretical insights can lead to actionable strategies for CLV prediction and subscription services. From segmentation and targeting strategies to pricing optimization and customer retention initiatives, theoretical frameworks serve as a blueprint for businesses to navigate the intricacies of the subscription marketplace and maximize customer lifetime value.

2.4.1 Economic Value Theory And Subscription-based Businesses

Economic value theory is a foundational concept in understanding consumer behaviour and decision-making processes. It involves how individuals perceive and assess the value of products or services from their personal perspective. In the context of subscription-based services, economic value theory plays a critical role in dictating customer acquisition, retention, and ultimately, customer lifetime value. This review covers the complexities of economic value theory, exploring its fundamental principles, key components, and practical implications for subscription-based businesses. By examining factors such as utility, price sensitivity, and perceived benefits, this review seeks to explain how the economic value theory influences subscriber behaviour and guides strategic decision-making in subscription-based services.

The economic value theory centers around the school of thought that individuals make rational choices based on their perceptions of value. It asserts that consumers seek to maximize their utility, or satisfaction, derived from the consumption of goods and services, given their limited resources. This theory is built upon the fundamental principles of subjective perception of value, utility maximization, and marginal utility. In subjective perception of value, economic value is considered to vary from one individual to another based on their preferences, needs, and circumstances, implying that what one consumer considers as valuable may differ significantly from another. This underscores the importance of understanding consumer heterogeneity in value assessment which is one of the key concepts in modern-day CLV prediction approach. In utility maximization, consumers seek to maximize their utility by allocating their limited resources such as time and money to obtain goods and services that offer the highest level of satisfaction or the highest utility, emphasizing the importance of understanding how consumers perceive and evaluate the value proposition of products or services. Marginal utility principle asserts that as consumers consume more of a particular good or service, the additional satisfaction derived from each additional unit diminishes. The implication of the marginal utility principle is that businesses must balance pricing strategies with perceived value to maximize revenue and profitability.

2.4.2 Practical Implications Of Economic Value Theory For Subscription-based Businesses

Price sensitivity, perceived benefits, and opportunity costs are components of the economic value theory that influence consumers' perceptions and assessments of value. Price sensitivity which is also known as elasticity of demand, refers to consumers' responsiveness to price variations. It, in principle, mirrors the degree to which variations in price impact the quantity demanded of a product or service; thus, understanding price sensitivity is important for pricing strategies. This guides businesses on the optimal pricing levels that balance profitability with consumer affordability and perceived value. Perceived benefits play a decisive role in forming consumers' perceptions of value and willingness to pay for subscription services. These benefits may be functional attributes such as quality, performance, and reliability, emotional appeals such as brand image, status, and prestige, and experiential factors such as convenience, customization, and personalization. In opportunity costs, consumers evaluate the benefits of subscribing to a product or service against the benefits of alternative uses of their resources, such as purchasing competing products or services, investing in other activities. This insight helps businesses match their value proposition with consumers' preferences and priorities, boosting the attractiveness of their subscription services with respect to alternative options.

Economic value theory impacts diverse aspects of subscription-based businesses from subscriber acquisition, to retention, and CLV optimization. It provides the insights that guides businesses on developing strategic initiatives that align with consumers' preferences and promote the perceived value of subscription services. The implications range from pricing strategies, value communication, customer experience promotion, to data-driven decision-making. Pricing is critical to shaping consumers' perceptions of value and eagerness to subscribe to a service. Economic value theory provides that businesses should conform their pricing strategies to consumers' perceived benefits and price sensitivity regimes. This may involve offering tiered pricing plans that cater to different segments of customers based on their preferences and willingness to part with their resources. Furthermore, businesses can utilize dynamic pricing algorithms to optimize pricing in real-time with respect to variations in demand, competition, and other market dynamics. However, this is beyond the scope of this study. Effectively communicating the value proposition of subscription services is essential to attract and retain subscribers. Businesses must streamline the tangible and intangible benefits of their offerings, highlighting features such as exclusive content, personalized recommendations, and seamless user experiences. Usage of persuasive messaging techniques can be of great advantage, and together with social proof mechanisms, can promote consumers' perceptions of value and stimulate demand for subscription services. Promoting the subscriber experience is important for fostering long-term relationships and maximizing CLV. Economic value theory emphasizes the importance of delivering superior value relative to competitors, thereby incentivizing subscribers to remain loyal and continue their subscriptions. Investing in product innovation, service excellence, and customer engagement initiatives are various strategies that businesses can leverage to exceed consumers' expectations and boost their overall satisfaction. By leveraging advanced analytics and predictive modeling techniques, businesses can gain insights into consumers' preferences, behaviours, and lifetime value potential. This enables personalized marketing efforts, targeted retention strategies, and proactive churn prevention initiatives that promotes the economic value proposition for subscribers and drive sustainable growth for businesses.

2.4.3 Rationale And Relevance Of Economic Value Theory To The Study

The integration of economic value theory (EVT) within the context of this study is based on its theoretical merits and practical implications in guiding strategic decision-making in marketing analytics. EVT serves as a basic framework for understanding how consumers perceive value, make purchasing decisions, and accordingly allocate their resources. By involving the economic principles governing consumer behavior, the study can highlight the complexities of CLV prediction, which ultimately centers around subscribers' valuation of products or services over their lifetime. EVT offers insights into factors such as utility maximization, price sensitivity, and subjective perceptions of value, which are central to estimating and optimizing CLV. In addition, EVT offers a complete perspective that transcends traditional economic concepts, encompassing psychological, sociological, and behavioural dimensions of value perception. In the context of CLV prediction, this multipronged approach enables the study to capture the nuances of customer preferences, needs, and decision-making processes. With insights from EVT, this study can develop more robust predictive models that elucidate the diverse factors influencing CLV, thereby enhancing the accuracy and applicability of the findings. EVT also offers a standard framework for evaluating pricing strategies, marketing tactics, and subscriber engagement initiatives with respect to their impact on CLV. EVT offers practical implications for segmentation, targeting, and positioning strategies in CLV prediction. By segmenting subscribers based on their value perceptions, preferences, and behaviour patterns, businesses can target their offerings and marketing efforts to different subscriber segments effectively. This targeted approach enhances subscriber satisfaction, loyalty, and retention, ultimately driving CLV and fostering sustainable relationships with customers. By rooting the study in EVT principles, researchers can assess the effectiveness and efficiency of various strategies in maximizing subscriber value and long-term profitability. This analytical perspective helps businesses to prioritize investments, allocate resources strategically, and optimize their marketing mix to achieve sustainable growth and competitive advantage.

2.5 Conceptual Framework

2.5.1 Understanding the Role of Customer Demographics in CLV Prediction

Customer demographics play a critical role in customer lifetime value prediction. Demographic variables cover a broad range of characteristics that offer useful insights into the various profiles, preferences, and behaviours of subscribers within subscription-based services. In this review, age, gender, income level, and geographic location and how they influence CLV dynamics and inform strategic decision-making for businesses shall be examined.

Age is a basic demographic construct with critical implications for CLV prediction as diverse age categories are characterized with distinct preferences and behaviours. For instance, millennials, often profiled as being digital savvy with leaning towards experiential consumption, may show higher engagement levels with subscription-based services that provide personalized and interactive experiences, while the older generations may prioritize reliability, convenience and value for money, impacting their subscription retention rates and long-term loyalty. This implies that businesses need to understand these diverse preferences and therefore fashion their marketing and customer experience strategies to align with specific demographics if they must maximize their CLV. As passive as it may seem, gender impacts consumers' preferences, their purchase decisions, and the way they interact with certain brands. These differences are often deciphered through their interests and lifestyles. For example, subscription-based businesses offering beauty and fashion products may customer base dominated by the female gender as they are naturally inclined to self-care and cosmetic enhancements. This is in sharp contrast with the male folks who are more inclined to gaming, betting, and technology. Thus, having a grasp of gender-based preferences may well help businesses maximize their CLV through strategic product offerings. Income level is another construct and a critical determinant of purchasing power and discretionary spending, significantly impacting CLV dynamics. This implies that customers with higher income levels may be more likely to spend on top-tier levels of certain subscription services, a sharp contrast with to their low income counterparts who may prioritize affordability, value for money, and a mere access to basic necessities. Businesses must be able to strategise on the trade-off between affordability and value inclinations among their subscribers to maximize CLV. Geographic location may impact CLV on account of differing regional and cultural preferences and market dynamics. Beyond these, there are logistics considerations like shipping costs, delivery time, and availability, in addition to subscribers' lifestyle which may greatly be impacted by their locations like urban, suburban, metropolis.

Purchase behaviour metrics provide insights on the average order value, frequency, recency, and monetary value, and total lifetime spend of customer transactions. Leveraging historical transaction data, businesses can identify patterns and trends in customer purchasing behaviour, shaping predictive models and segmentation strategies. Recency, is a critical indicator that measures how responsive a subscriber is, in relation to recent purchases made. It is a measure that points to engagement, activeness, and loyalty and can be used by businesses to track retention and implement personalized strategies. It is important to note that recency is a major input to the BG/NBD and Pareto/NBD models, in consonance with temporal dynamics to forecast future transactions and ultimately CLV; thus, imputing recency into CLV prediction models enables businesses to identify the time-sensitive nature of customer interactions and fashion their retention efforts accordingly. How often a customer makes purchases within a given timeframe is what is referred to as frequency and it offers insights into customer loyalty, engagement, and purchasing habits. A higher level of commitment, brand loyalty is exhibited when customers make frequent purchases. Frequency is crucial to predictive model, hinting businesses of future purchase behaviour, customer needs and preferences. Monetary value is a measure of the amount spent by a subscriber on purchases, offering insights into customer spending profiles, preferences, and purchasing power. The higher the amount spent on purchases or the higher the average order value the greater the overall revenue and ultimately the higher the CLV. Total lifetime spend is another critical component to CLV prediction. It is the cumulative amount spent by subscribers over their entire relationship with the business. Long-term loyalty and value to the business are exhibited by subscribers with higher total lifetime spend. Analyzing total lifetime spend helps businesses to recognise valuable customers, implement targeted retention strategies, and cultivate long-term relationships that drive sustainable growth. Customer engagement metrics measures the level of interaction and involvement of subscribers with the service platform. Duration of subscriptions, number of referrals, and frequency of platform visits are indicative of subscribers’ engagement. High levels of customer engagement lead to increased CLV and long-term profitability. Product/service usage metrics serve as a critical construct that help businesses gain insights into the extent to which subscribers interact with the offerings of the subscription service. Metrics such as frequency of product usage, feature adoption rate, and customer feedback and satisfaction scores can be used by businesses to gauge the value recommendation thereby shaping shaping CLV dynamics and driving strategic decision-making for businesses.

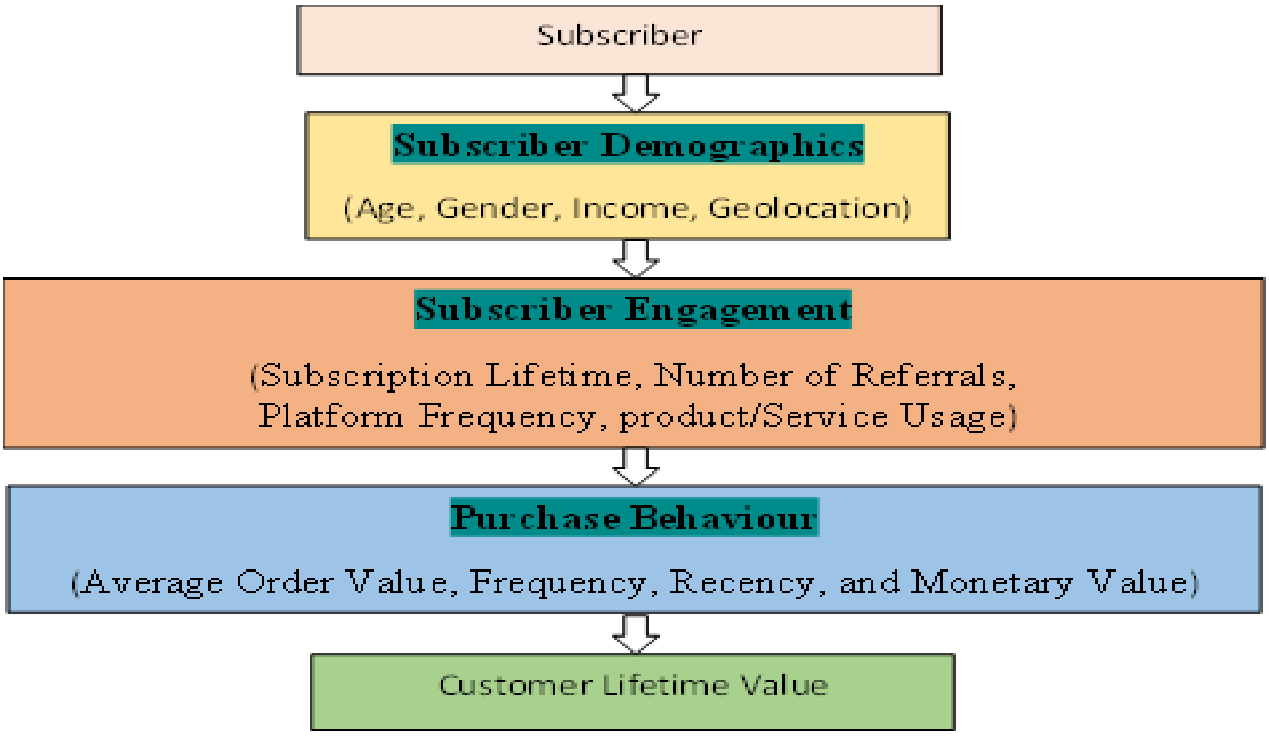

Subscriber
Subscriber Demographics
(Age, Gender, Income, Geolocation)
Subscriber Engagement
(Subscription Lifetime, Number of Referrals, Platform Frequency, product/Service Usage)
Purchase Behaviour
(Average Order Value, Frequency, Recency, and Monetary Value)
Customer Lifetime Value

Figure 1: Conceptual framework for the study (by the Researcher)

2.6 Research Gap

Despite the extensive research carried out on customer lifetime value prediction models, there are several critical gaps that militate against the development of accurate and scalable models, particularly within subscription-based businesses. A comprehensive investigation of the existing literature reveals three significant research gaps that warrant further investigation, namely failure to recognize the probabilistic nature of subscribers' interactions in CLV modeling, non-inclusion of customer engagement metrics, and economic value theory when modeling CLV. Many CLV models miss out on accurate prediction of customer behaviour which have direct bearing on long-term loyalty, retention and ultimately profitability as a result of ignoring the criticality of engagement metrics. Subscriber interactions within subscription-based businesses are stochastic in nature due to uncertainty in customer behaviour, different engagement levels, and even their networking profile. This is where many existing CLV models fail because they are unable to adequately account for these probabilistic elements, leading to reduced predictive accuracy and limited applicability in real-world scenarios. For instance, the rate at which subscribers discontinue their subscriptions cannot be accurately predicted with deterministic models due to uncertainties associated with customer behaviours. Customer engagement levels which are also unstable over time and are impacted by diverse external factors, further contribute to the probabilistic nature of subscriber interactions. Limited integration of economic value theory into existing models also makes other models unreliable. Economic value theory examines how consumers apportion value to different goods and services based on their utility and preferences; thus, factoring this theory is critical but this is what is overlooked by most researchers. Economic value theory establishes that subscribers' feedback, reviews, and satisfaction scores are as crucial as their transactional interactions. Addressing this research gap requires the development of more robust CLV prediction models that explicitly account for the probabilistic nature of subscriber interactions. Probabilistic models, such as Markov Chain analysis and Pareto/NBD approaches offer promising approaches for incorporating uncertainty into CLV prediction models. Furthermore, integrating advanced analytical techniques, such as machine learning algorithms and data-driven simulations, can boost the predictive accuracy of CLV models by capturing intricate patterns and interactions within subscription-based datasets. By accepting a stochastic approach to CLV prediction, researchers can develop models that are better equipped to address the uncertainties associated with subscription-based businesses, thereby enhancing their usage in real-world applications.

2.7 Critical Analysis of Findings

The journey of Customer Lifetime Value (CLV) metrics has witnessed a profound transformation from its origins as a basic financial measure to its current status as a sophisticated tool for understanding and enhancing subscriber relationships. Initially, CLV was viewed merely as a financial metric indicating the projected revenue from a customer over their lifetime. However, recent advancements have expanded this perspective, bringing together strategic and relational dimensions that reflect the evolving nature of subscriber interactions and business objectives. This progression in CLV prediction models has been characterized by significant improvements in accuracy and sophistication. Early models such as RFM (Recency, Frequency, Monetary) were foundational but limited in their ability to capture the complexities of customer behavior. Modern approaches, including Pareto/NBD (Negative Binomial Distribution) and Markov Chains, have introduced advanced methodologies that account for probabilistic customer behavior and temporal dynamics. These models offer a more streamlined understanding of customer value by incorporating various factors like transaction frequency, purchase recency, and customer retention patterns which this study considers as state transitions.

The rise of subscription-based business models has further influenced the landscape of CLV prediction. These predictive models, characterized by recurring revenue and continuous customer engagement, highlight the importance of personalized interactions and long-term customer relationships. The subscription model's emphasis on ongoing value delivery aligns with the shift towards customer-centric strategies aimed at maximizing CLV. This is where Economic Value Theory (EVT) plays a crucial role in shaping contemporary CLV models. EVT emphasizes the subjective nature of value perception and utility maximization, providing a theoretical framework that helps businesses understand consumer decision-making processes. This theoretical underpinning is essential for developing marketing strategies that align with consumer expectations and enhance CLV. Despite these advancements, challenges remain in the application of predictive models. The complexity of advanced models, along with their underlying assumptions, can sometimes create gaps between theoretical predictions and real-world subscriber behaviour. This underscores the need for businesses to carefully adapt and validate these models to ensure their effectiveness in practical scenarios.

2.8 Implications of the Findings for this Research

The insights and critical analysis derived from the literature review provide a robust foundation for this research on CLV prediction. The evolution of CLV metrics and the advancements in predictive models highlight the importance of adopting a comprehensive approach to subscriber value measurement. By integrating both transactional and relational aspects, this research seeks to offer a more streamlined understanding of CLV which will facilitate the development of more effective marketing strategies by putting forward sophisticated methodologies for model development, implementation, and a detailed analysis of subscriber behaviour and retention patterns; providing valuable insights for optimizing marketing strategies and resource allocation. This premise forms the basis on which the Economic Value Theory (EVT) is incorporated into the analysis, to align the research with consumer value perceptions and decision-making processes. This theoretical foundation will support the development of strategies that resonate with subscriber expectations, ultimately enhancing CLV and fostering long-term relationships. By integrating advanced predictive models, applying EVT principles, and addressing the complexities of subscription-based models, this study will contribute to a deeper understanding of CLV and its implications for marketing strategies, ultimately leading to more effective and subscriber-centric approaches in business.

CHAPTER THREE: METHODOLOGY AND RESEARCH DESIGN

3.1 Introduction

The methodology and research design for the study covers several key stages, each crucial for the successful development and implementation of the predictive model for customer lifetime value. The first step in the research methodology involved collecting and preparing data for the analysis. This included gathering relevant datasets that covered customers' subscriptions, transactions, location, products or services, and demographics. This chapter seeks to explore the complexities of designing CLV estimation model for subscription-based businesses, critically assessing and juxtaposing the effectiveness of the developed predictive model which covers the RFM (Recency, Frequency, and Monetary) model, Markov Chain Analysis, and Pareto/NBD model. Through a thorough examination of the research approach, including data collection and preparation, model development, integration and comparison, model evaluation, implementation, documentation, and continuous improvement, this chapter establishes a detailed roadmap for understanding and harnessing the potential of the developed model for the estimation of CLV in subscription-based ecosystems. Beginning with meticulous data collection and preparation, the chapter delves into the terrain of model development, integration, and comparison, resulting in a comprehensive assessment of model performance and real-world implementation. With continuous improvement as the guiding principle, the research iteratively finetunes its methodologies, leveraging insights obtained from stakeholder feedback and evolving business landscapes.

The choice of RFM Analysis was based on the fact that it is well-suited for initial customer segmentation and understanding basic behaviour patterns of the subscribers. Given the dataset's attributes, RFM was employed to quickly provide insights into high-value customers and those at risk of churn. For this study, the RFM model effectively segmented subscribers based on their past purchasing behaviour, with segments such as 'Potential Loyalists,' 'Watch List,' and 'Loyal Subscribers' clearly identified and providing actionable insights into subscriber engagement and value. For example, the high monetary value and frequency in the 'Potential Loyalists' segment indicated significant revenue potential, while the 'Emerging Subscribers' segment highlights opportunities for growth with lower spending but higher frequency. However, while RFM is straightforward and requires minimal data preprocessing, making it easy to implement with the available dataset. It lacks the predictive power to forecast long-term CLV or changes in subscriber behaviour. This is where the BG/NBD and Pareto/NBD models came handy. BG/NBD and Pareto/NBD Models offer more advanced probabilistic approaches for predicting CLV. They were particularly chosen for their usefulness and alignment with datasets having detailed transactional data, like Monthly Revenue and pay_date columns in the Netflix data. Both models have the reputation of handling varying purchase frequencies and dropout rates, providing robust CLV predictions; thus, the choice between BG/NBD and Pareto/NBD would depend on the specific data distribution and the complexity of parameter estimation. However, while BG/NBD and Pareto/NBD are robust in predicting future value of subscribers, they still lack the ability to uncover the transitions that play our as subscribers engage at varying frequencies. Most literatures prefer to discuss extremities, for example, “active” and “churn”, ignoring what really transpires in-between. The purpose of this study was to uncover these transitions, whether at steady-state or in transiency, and integrate the insights so obtained into the overall predictive model. In other words, this study considered that it is not enough to predict future value but that the ability to anticipate the behaviour of the subscribers was also critical.

3.2 Model Development and Methodology

The research approach for this study adopts a quantitative methodology to explore CLV dynamics within subscription-based enterprises leveraging statistical techniques and predictive modeling and critically examining customer transaction data and its underlying patterns, in order to forecast CLV with precision. The statistical techniques employed in this study are to help dissect and analyze customer transaction data meticulously curated and aggregated to ensure the validity and reliability of the subsequent analyses and findings. These techniques include exploratory data analysis, descriptive and inferential statistics, to offer a systematic framework for uncovering patterns, trends, and relationships within the data. Descriptive statistics provide a snapshot of key metrics such as mean transaction value, frequency of purchases, and average order value, offering valuable insights into the overall performance of the services, while inferential statistics will aid in drawing inferences and making predictions about broader landscape of subscription-based businesses based on the available sample data, laying the foundation for predictive modeling. The predictive modeling process shall include data preprocessing, model selection, training, validation, and evaluation. Through iterative finetuning and optimization, this model will capture the underlying dynamics of CLV within subscription-based enterprises.

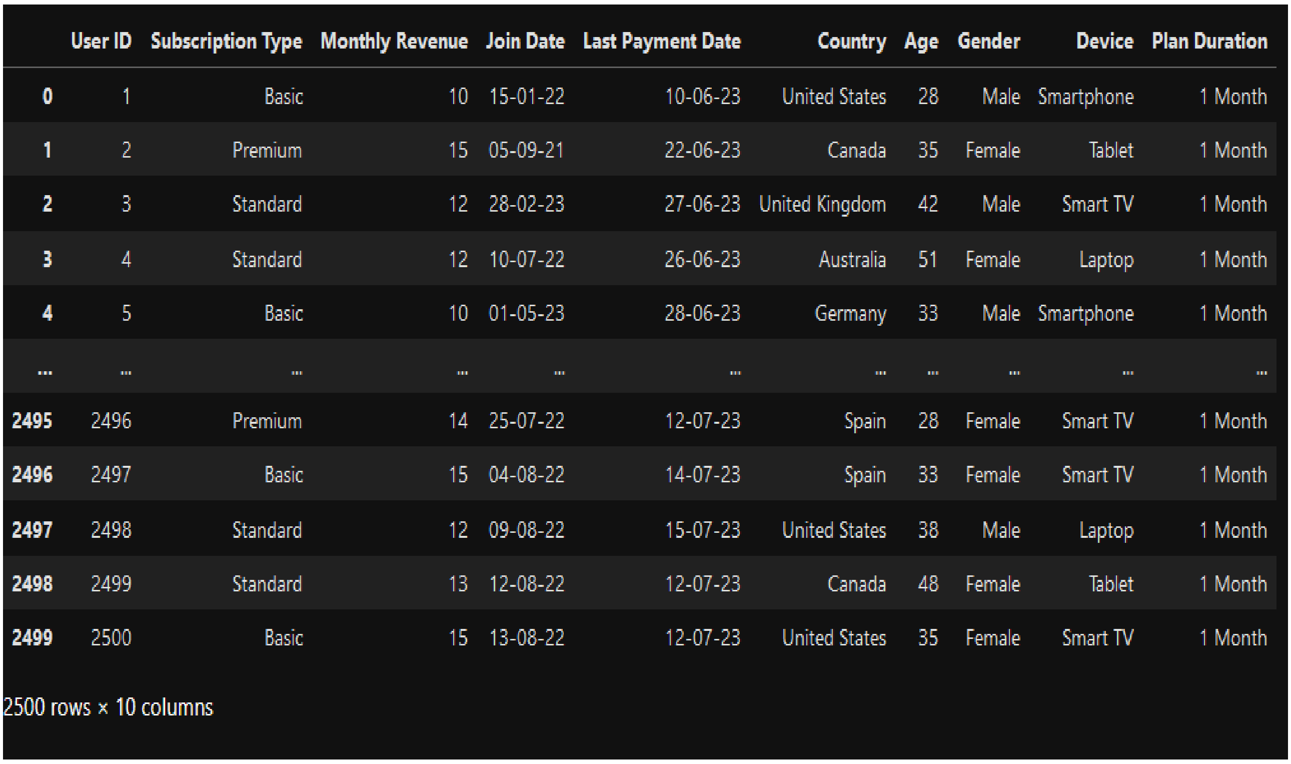
The process of data collection and preparation for this study is of immense significance and critical to ensuring the validity and reliability of subsequent analyses. Deducing from the work of Bishop and Boyle (2019), and Surucu and Maslakci (2020), data validity has to do with the data meeting purpose for which the data was collected in the first place, while data reliability refers to consistency of data each time the analysis was repeated; thus, meticulous data collection and preprocessing are essential for accurate CLV prediction. The first step involves gathering, cleaning, and preprocessing for analysis relevant datasets that cover users’ subscriptions, transactions, location, products or services, and demographics. Preprocessing is carried out to remove errors, outliers, inconsistencies, and missing values from the datasets. Then the data is formatted in a standardized manner suitable for the analytic purposes. Clean and standardized data ensure the robustness and reliability of subsequent analyses and model development. This is immediately followed by feature engineering, a process that involves creating additional but relevant features, removing extraneous features or transforming existing ones to improve model performance. The goal is to ensure that only necessary features are available for accurate CLV prediction (Kuhn and Johnson, 2020). RFM analysis, churn rate calculation, and engagement metrics generation are applied to extract meaningful insights from the data; thus, effective feature engineering boosts the predictive power of the model by capturing essential characteristics and behaviours of subscribers that impact CLV. For example, for the RFM analysis, subscribers will be segmented based on their recency, frequency, and monetary value of transactions, to identify high-value segments. So, the features would include join date, pay date and the amount, payment frequency and recency will be necessary features for accurate classification as was established by Rawat and Khemchandani (2017). The subscription data will be leveraged to model customer subscription patterns using techniques such as Pareto/NBD modeling, while Markov Chain will be leveraged for state transition analysis to highlight the dynamics of customer engagement over time.

The study leverages secondary data collection from an online resource, Netflix subscribers database. Netflix is an American subscription streaming service that provides a library of movies, TV shows, and original content to its members. It operates a subscription model where users pay a monthly fee to access the entire content library. There are typically different subscription tiers offering varying levels of streaming quality, example, standard definition and high definition or the number of devices that can stream simultaneously.

Table 3.2a gives a snapshot of the subscription data table which contains 10 columns. The country column contains 10 countries, while the recency of the data points ranges from 2021 to 2023. The subscription type comprises 3 distinct categories, namely basic, standard, and premium.

CLV metrics provides insights into different aspects of customer value and profitability. The data table contains a range of features that influence CLV dynamics and customer behaviours within subscription services. These variables include customer demographic variables such as age, gender, and geographic location provide insights into the socio-economic characteristics of the subscribers and their purchasing preferences. Subscription plan type is another variable of interest, and this includes basic, standard, and premium subscriptions. Other variables are recency, frequency, monetary, RFM score which will be derived from the dataset and will also be used to segment the customers.

Table 3.2a: A snapshot of the Netflix subscription data table



	User ID	Subscription Type	Monthly Revenue	Join Date	Last Payment Date	Country	Age	Gender	Device	Plan Duration
0	1	Basic	10	15-01-22	10-06-23	United States	28	Male	Smartphone	1 Month
1	2	Premium	15	05-09-21	22-06-23	Canada	35	Female	Tablet	1 Month
2	3	Standard	12	28-02-23	27-06-23	United Kingdom	42	Male	Smart TV	1 Month
3	4	Standard	12	10-07-22	26-06-23	Australia	51	Female	Laptop	1 Month
4	5	Basic	10	01-05-23	28-06-23	Germany	33	Male	Smartphone	1 Month
...	...	...	...	...	...	...	...	...	...	...
2495	2496	Premium	14	25-07-22	12-07-23	Spain	28	Female	Smart TV	1 Month
2496	2497	Basic	15	04-08-22	14-07-23	Spain	33	Female	Smart TV	1 Month
2497	2498	Standard	12	09-08-22	15-07-23	United States	38	Male	Laptop	1 Month
2498	2499	Standard	13	12-08-22	12-07-23	Canada	48	Female	Tablet	1 Month
2499	2500	Basic	15	13-08-22	12-07-23	United States	35	Female	Smart TV	1 Month

2500 rows × 10 columns

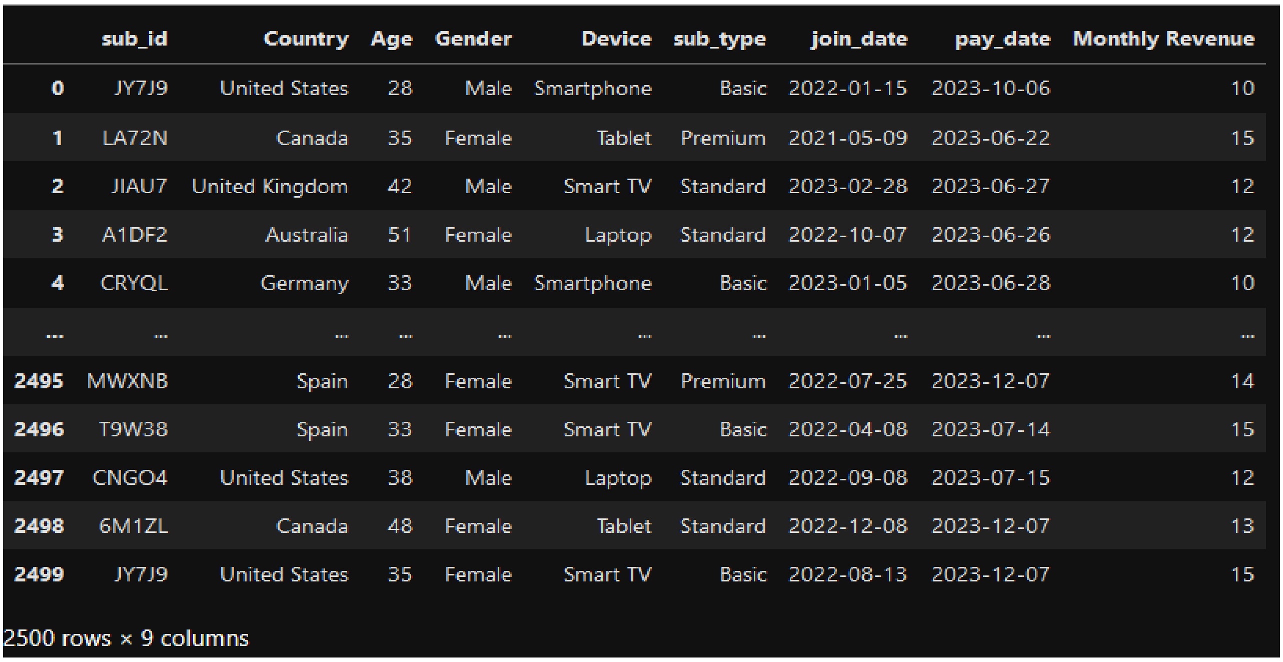
Table 3.2b: Data dictionary for the Netflix data table

Column Name	Description	Data Type
User ID	Unique identifier for each user	Integer
Subscription Type	Type of subscription plan	Categorical
Monthly Revenue	Revenue generated from the subscription per month	Numeric
Join Date	Date when the user joined the subscription	Date
Last Payment Date	Date of the most recent payment	Date
Country	Country of residence of the user	Categorical
Age	Age of the user	Integer
Gender	Gender of the user	Categorical
Device	Device used by the user for the subscription	Categorical
Plan Duration	Duration of the subscription plan	Categorical

Data Transformation, Feature Engineering and RFM

The first step in preparing the dataset involved converting date columns to a datetime format and removing unnecessary columns. This ensured that the data was in the correct format and streamlined for analysis. The dataset was later transformed into a much more suitable form that can easily allow for derivation of Recency, Frequency, and Monetary features. On inspection, the user id was considered to be inappropriate for the task ahead; thus, a better unique identifier called sub_id was derived using the country and subscription type columns. Reason for this was that the countries would be anonymized, while also stratifying the subscribers based on subscription types; thus, each country was accurately represented across the different subscription types. This resulted in Table 3.2a.

Table 3.2c: Transformed table for the RFM derivation



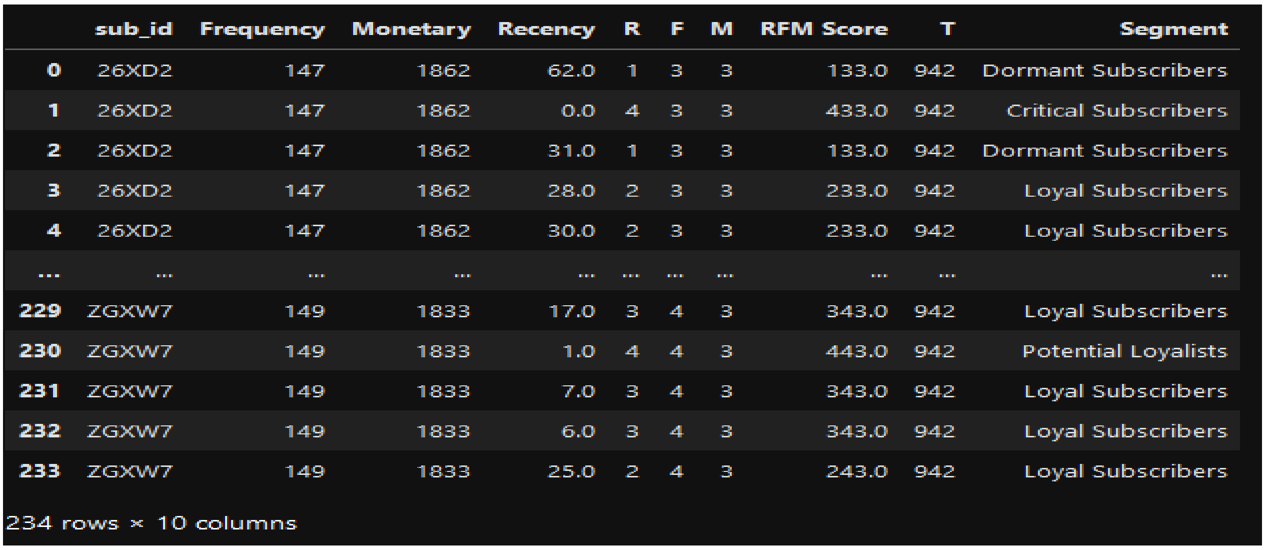
	sub_id	Country	Age	Gender	Device	sub_type	join_date	pay_date	Monthly Revenue
0	JY7J9	United States	28	Male	Smartphone	Basic	2022-01-15	2023-10-06	10
1	LA72N	Canada	35	Female	Tablet	Premium	2021-05-09	2023-06-22	15
2	JIAU7	United Kingdom	42	Male	Smart TV	Standard	2023-02-28	2023-06-27	12
3	A1DF2	Australia	51	Female	Laptop	Standard	2022-10-07	2023-06-26	12
4	CRYQL	Germany	33	Male	Smartphone	Basic	2023-01-05	2023-06-28	10
...	...	...	...	...	...	...	...	...	...
2495	MWXNB	Spain	28	Female	Smart TV	Premium	2022-07-25	2023-12-07	14
2496	T9W38	Spain	33	Female	Smart TV	Basic	2022-04-08	2023-07-14	15
2497	CNGO4	United States	38	Male	Laptop	Standard	2022-09-08	2023-07-15	12
2498	6M1ZL	Canada	48	Female	Tablet	Standard	2022-12-08	2023-12-07	13
2499	JY7J9	United States	35	Female	Smart TV	Basic	2022-08-13	2023-12-07	15

2500 rows × 9 columns

The RFM values in this study are crucial for customer segmentation and CLV estimation. Recency is calculated as the number of days between consecutive payments, Frequency as the number of transactions a customer has made, and Monetary as the total revenue generated from a customer. This step involved sorting the data by subscriber ID and payment date to calculate these metrics accurately. The resulting RFM values were used to score customers based on their behaviour, facilitating the segmentation process.

The RFM scores were assigned using quartiles, and customers were segmented based on their RFM scores. Frequency and Monetary were rated on the same scale, while Recency was rated on a reverse scale.

Table 3.2d: RFM segmentation table



	sub_id	Frequency	Monetary	Recency	R	F	M	RFM Score	T	Segment
0	26XD2	147	1862	62.0	1	3	3	133.0	942	Dormant Subscribers
1	26XD2	147	1862	0.0	4	3	3	433.0	942	Critical Subscribers
2	26XD2	147	1862	31.0	1	3	3	133.0	942	Dormant Subscribers
3	26XD2	147	1862	28.0	2	3	3	233.0	942	Loyal Subscribers
4	26XD2	147	1862	30.0	2	3	3	233.0	942	Loyal Subscribers
...	...	...	...	...	...	...	...	...	...	...
229	ZGXW7	149	1833	17.0	3	4	3	343.0	942	Loyal Subscribers
230	ZGXW7	149	1833	1.0	4	4	3	443.0	942	Potential Loyalists
231	ZGXW7	149	1833	7.0	3	4	3	343.0	942	Loyal Subscribers
232	ZGXW7	149	1833	6.0	3	4	3	343.0	942	Loyal Subscribers
233	ZGXW7	149	1833	25.0	2	4	3	243.0	942	Loyal Subscribers

234 rows × 10 columns

The following is the frequency and monetary (FM) scale

FM Scale:	4	3	2	1

While for

Recency Scale:	1	2	3	4

Thus, this logic formed the basis for the segmentation as further elucidated by this combination:

Elite Subscribers: High Frequency (4) and Monetary (4) with the most recent engagement (1).

Loyal Subscribers: Moderate Recency (2, 3) with high Frequency (3, 4) and Monetary (3, 4).

Potential Loyalists: Recent Recency (1, 2) with moderate Frequency (2, 3) and Monetary (2, 3).

First-Time Subscribers: Most recent engagement (1) with low Frequency (1) but any Monetary value (1, 2, 3, 4).

Emerging Subscribers: Recent Recency (1, 2) with low Frequency (1) and Monetary (1, 2).

Watch List: Older Recency (3, 4) with moderate Frequency (2, 3) and Monetary (2, 3).

Dormant Subscribers: Least recent engagement (4) with high Frequency (4) and Monetary (4).

Critical Subscribers: Least recent engagement (4) with high Frequency (4) and moderate Monetary (3).

Fading Subscribers: Older Recency (3, 4) with low Frequency (1) and low Monetary (1, 2).

Lost Subscribers: Least recent engagement (4) with low Frequency (1) and low Monetary (1).

Others: Any other combination that doesn't fit into the above categories.

Table 3.2e: Table displaying how the segments were generated using the RFM Score

R	F	M	Score	Segment
1	4	4	144	Elite Subscribers
2	3	3	233	Loyal Subscribers
1	2	3	123	Potential Loyalists
1	1	4	114	First-Time Subscribers
2	1	2	212	Emerging Subscribers
3	2	3	323	Watch List
4	4	4	444	Dormant Subscribers
4	4	3	443	Critical Subscribers
3	1	2	312	Fading Subscribers
4	1	1	411	Lost Subscribers
3	2	1	321	Others

The RFM dataframe obtained was subsequently transformed into a summary table. This summary was adopted to arrive at robuts performance and to ensure that all the customers were given equal chance. This summarized the subscription data table into 27 distinct subscribers. The emerging dataframe was then partitioned into 3, namely training, validation, and testing. This partitioning was done once and globally for the individual models, namely BG/NBD and Pareto/NBD to copy from using the copy() function. The BG/NBD was coded as bgf, while Pareto/NBD was coded as pareto. Thus, the training data for the bgf model was copied from the train.copy() and assigned to bgf_train, validation.copy() was assigned tobgf_validation, while test.copy() was assigned to bgf_test. The penalizer coefficient, penalizer_coef was set at 10. The bgf model was fitted with the frequency (F), recency (R) and age (T) columns of the bgf_train dataframe. The T represents the specific period or age of the transaction in days, taking 30 days as the standard number of days for a month. After running the code, a summary dataframe for the bgf model was obtained. The parameters on the table include coef, se(coef), lower 95% bound, upper 95% bound on the columns, while the rows have the r, alpha, a, and b. The validation of the model was performed with bgf_validation dataframe where bgf.conditional_expected_number_of_purchases_up_to_time(), the bgf customer lifetime value prediction module was called on the Monetary, T, Frequency, and Recency. The test data, bgf_test dataframe was used to test the model for a period of 182.5 days (6 months) and the results were noted. The evaluation metrics used for the model were mean absolute error (MAE), mean standard error (MSE), and root mean standard error (RMSE).

For the pareto model, the training data for was also copied from the train.copy() and assigned to pareto_train, validation.copy() was assigned to pareto_validation, while test.copy() was assigned to pareto_test. The penalizer coefficient, penalizer_coef was also set at 10. Just like the bgf model, it was fitted with the frequency (F), recency (R) and age (T) columns of the pareto_train dataframe. The T represents the specific period or age of the transaction in days, taking 30 days as the standard number of days for a month. The summary dataframe for the pareto model was slightly adjusted, thus, the output did not show up like that of the bgf model on account of the fact that the pareto model lacks parameters like the se(coef), lower 95% bound, upper 95% bound. The adjustment for pareto was using the pareto.fitter.params_ method. Thus, the summary table has only one column called coef but same number of row parameters: r, alpha, a, and b. The validation of the model was performed with pareto_validation dataframe and tested with pareto_test dataframe. Like the bgf model, the function was called on the Monetary, T, Frequency, and Recency fields for a period of 6 months using the pareto_fitter.conditional_expected_number_of_purchases_up_to_time(). Just like the evaluation metrics used for the bgf model, the mean absolute error (MAE), mean standard error (MSE), and root mean standard error (RMSE) evaluation metrics were also used for the pareto model. The both models are built into Dataframes that contains the sub_id, predicted purchase quantities, and predicted customer lifetime values.

From the RFM table, a transition matrix was derived. The rule is that the content of the matrix table must sum up to 1 which is in order because probability ranges between 0 and 1, where 1 signifies that an event will occur and 0 a probability that an event will not occur. The transition displays the probabilities of subscribers moving from one state to another which in this case are the segments. For this study, there were 9 segments, namely “Potential Loyalists”, “Watch List”, “Loyal Subscribers”, “First-Time Subscribers”, “Fading Subscribers”, “Emerging Subscribers”, “Critical Subscribers”, “Elite Subscribers”, “Dormant Subscribers”. From the transition matrix, steady-state distribution can be computed and future predictions made. The predictions are about how the subscribers are likely to transit from one segment to another. This is more than merely looking out for churn as found in a good number of literatures. This is critical because it helps to track a subscriber’s movement before a possible eventual churn actually takes place. The steps for the prediction are 5, 10, 20, 50, and 100. The transition calculation generates a matrix that also estimates possible customer lifetime value across the different segments

The transition matrix is of this form:

$$\begin{bmatrix} P11 & \cdots & P19 \\ \vdots & \ddots & \vdots \\ P91 & \cdots & P99 \end{bmatrix}$$

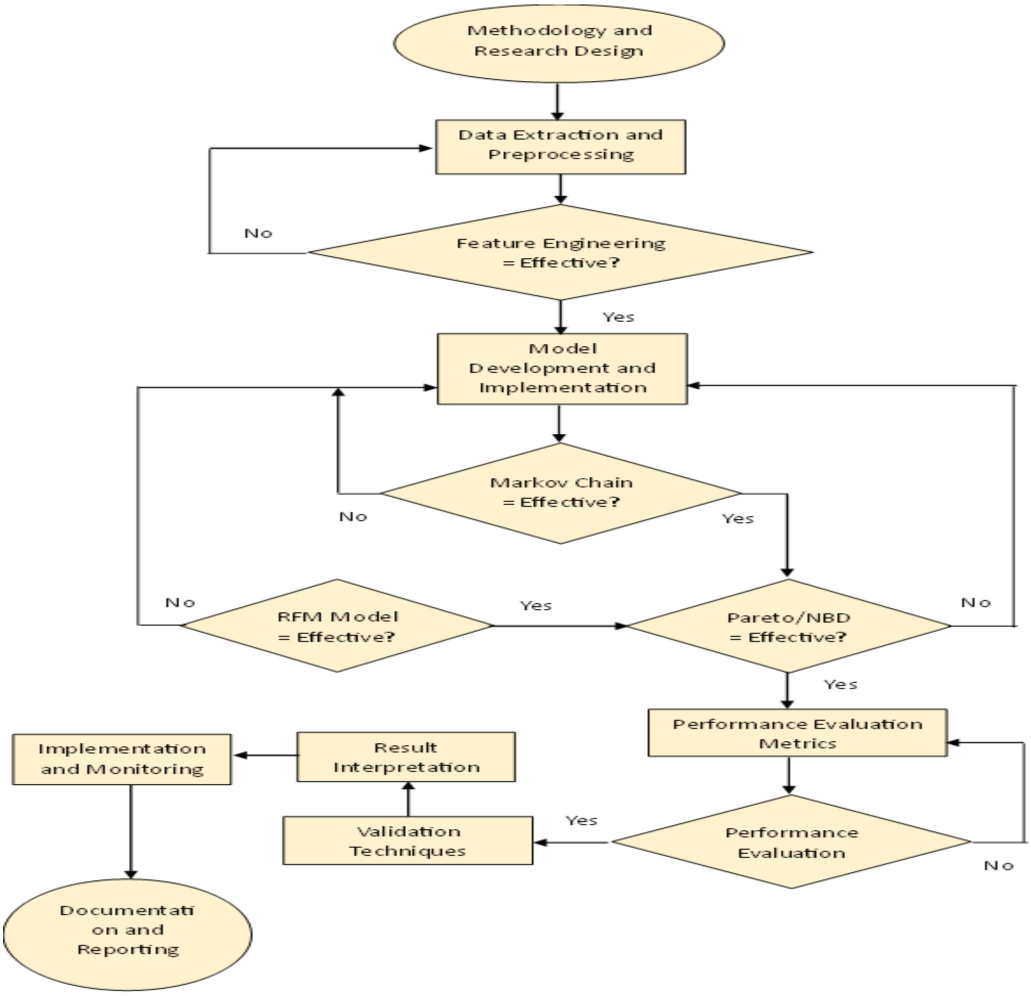

Methodology and Research Design
Data Extraction and Preprocessing
No
Feature Engineering = Effective?
Yes
Model Development and Implementation
Markov Chain = Effective?
No
Yes
No
RFM Model = Effective?
Yes
Pareto/NBD = Effective?
No
Yes
Performance Evaluation Metrics
Implementation and Monitoring
Result Interpretation
Validation Techniques
Yes
Performance Evaluation
No
Documentation and Reporting

Figure 2: Methodology flowchart for the model development (by the Researcher)

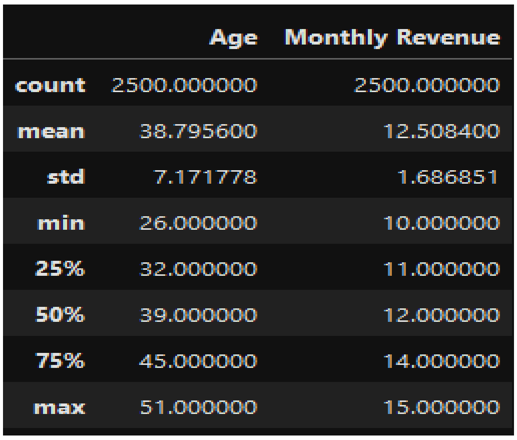
CHAPTER FOUR

MODEL IMPLEMENTATION AND CRITICAL EVALUATION

4.1 Model Development

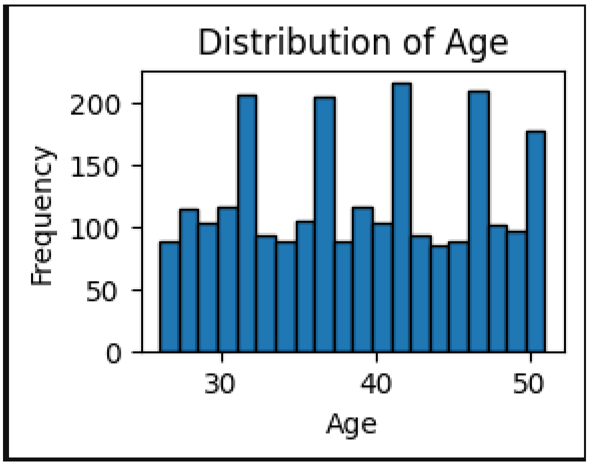
Understanding the age distribution and monthly revenue of subscribers was crucial in analyzing and predicting Customer Lifetime Value (CLV), hence, the usage of age as a construct. These metrics not only influence the segmentation of customers but also provide insights into how different demographic and revenue factors impact the overall value contributed by each subscriber. The age distribution of subscribers is a significant factor in understanding customer behavior and predicting CLV. With a total of 2,500 entries in the dataset, the age column has no missing values, ensuring a comprehensive analysis of subscriber demographics. The minimum age is 26 years, and the maximum is 51 years. This range indicates that the subscriber base includes both younger and older individuals. The average age of subscribers was 38.80 years. This central tendency suggests that the bulk of the customer base falls within the middle-aged group, which may indicate a stable and possibly more financially secure segment. The standard deviation of 7.17 years reflects variability in the age of subscribers suggesting that there is a diverse age range among subscribers, potentially affecting purchasing patterns and preferences. The 25th percentile at 32 years and the 75th percentile at 45 years provide a clearer picture of the distribution. This range indicates that the majority of subscribers fall between early 30s and mid-40s, with 25% of subscribers being younger than 32 years and 25% older than 45 years. The median age of 39 years, which is very close to the mean, further confirms that the age distribution is fairly symmetric with a central tendency around this age. understanding age distribution helps in segmenting customers effectively, for example, older subscribers (closer to 51 years) might have different purchasing behaviours and preferences compared to younger subscribers (around 26 years).

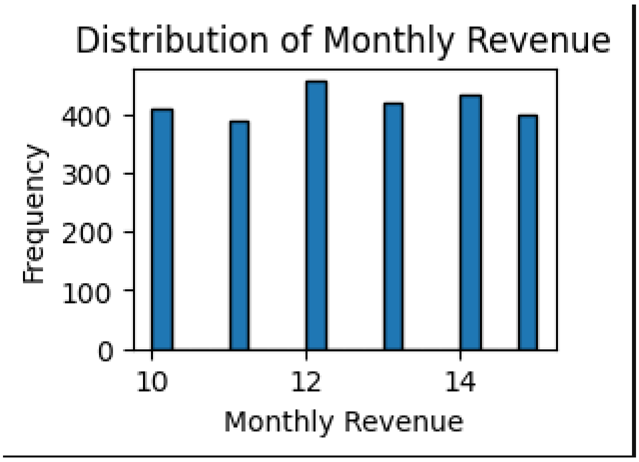
Table 4.1: Summary statistics



	Age	Monthly Revenue
count	2500.000000	2500.000000
mean	38.795600	12.508400
std	7.171778	1.686851
min	26.000000	10.000000
25%	32.000000	11.000000
50%	39.000000	12.000000
75%	45.000000	14.000000
max	51.000000	15.000000

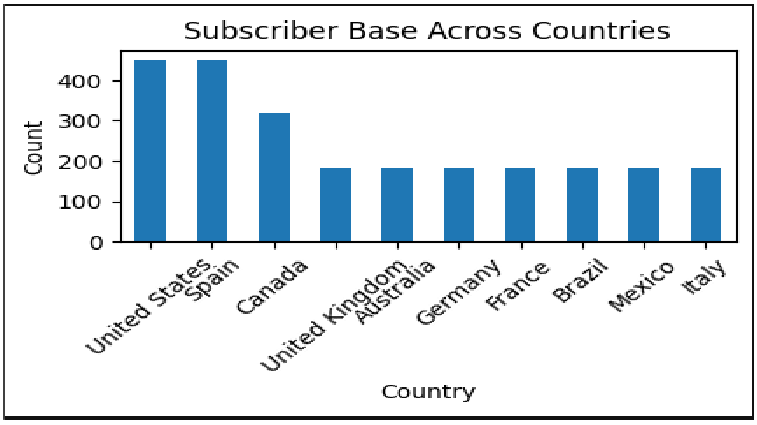
Older subscribers may exhibit more stable purchasing patterns, while younger ones might show higher variability. The average monthly revenue is $12.51, which indicates the typical revenue contribution per subscriber. This average is a useful benchmark for understanding the overall revenue generation potential of the subscriber base. With a standard deviation of $1.69, there is notable variation in the revenue amounts, an indication that while most subscribers contribute around the average, there are some with significantly higher or lower revenues. The minimum revenue recorded is $10, and the maximum is $15. This range reflects the lowest and highest revenue contributions from subscribers. The 25th percentile at $11 and the 75th percentile at $14 show that 50% of subscribers contribute between $11 and $14. This interquartile range provides a clearer view of where the bulk of revenue contributions lie, excluding the outliers at the extremes. The median revenue of $12 aligns closely with the mean, indicating that the revenue distribution is relatively symmetrical. This means that the majority of subscribers contribute around the average revenue, with few extreme outliers. Thus, subscribers contributing closer to the maximum revenue ($15) represent a higher value segment, whereas those at the minimum ($10) are less valuable. Identifying these revenue segments can help in developing strategies to maximize CLV, such as offering premium services to high-revenue subscribers or targeting low-revenue segments with upselling opportunities. This is because age and revenue are key variables that influence purchasing behaviour and loyalty.

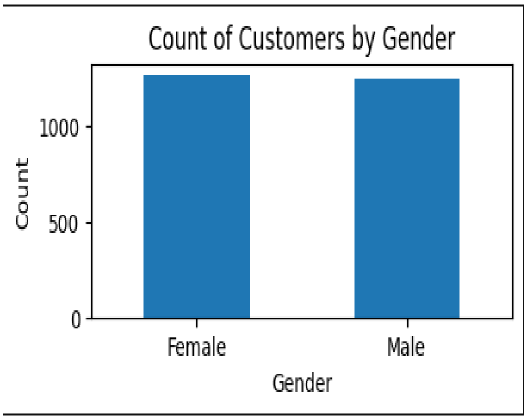




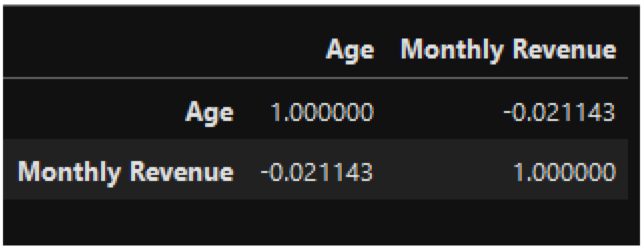
United States and Spain have the highest number of subscribers, with counts exceeding 400 each, Canada follows with a subscriber count slightly above 300, while other countries have fewer subscribers, each with counts less than 200. The high subscriber counts in the United States and Spain indicate these markets are particularly significant and higher subscriber counts in these countries can significantly impact overall revenue and CLV, while countries with fewer subscribers would require targeted promotional campaigns or localized offerings. Lower subscriber counts in some countries might indicate potential opportunities for market expansion. Identifying barriers to growth and developing strategies to overcome them could be beneficial.

The bar chart shows a nearly equal distribution of subscribers between male and female genders. Since the distribution is nearly equal, gender-based segmentation may not show significant differences in behaviour or CLV between male and female subscribers. However, it's still valuable for understanding any nuanced differences in preferences or purchasing patterns.





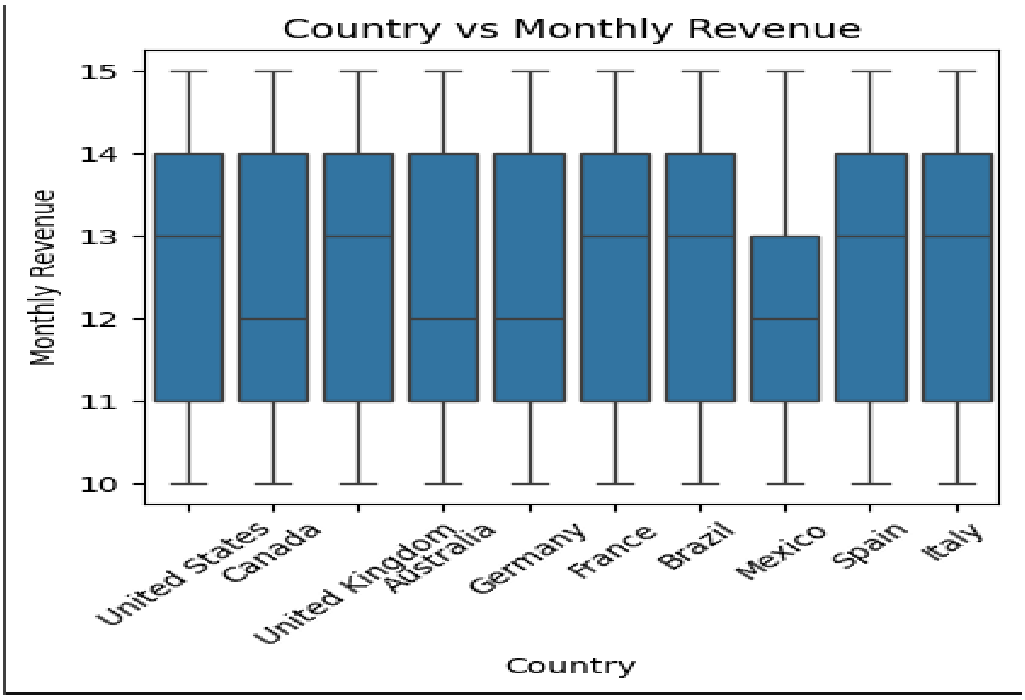
With a balanced gender distribution, marketing efforts can be designed to cater to both genders equally or tailored based on specific insights gathered from further analysis. The correlation value of -0.021 indicates a very weak negative relationship between age and monthly revenue. In practical terms, this suggests that age has little to no impact on monthly revenue, based on this dataset. Furthermore, monthly revenue appears relatively stable across different age groups, indicating that age is not a significant factor in determining how much subscribers spend monthly. Other features, such as subscription type or country, might be more relevant for predicting CLV.



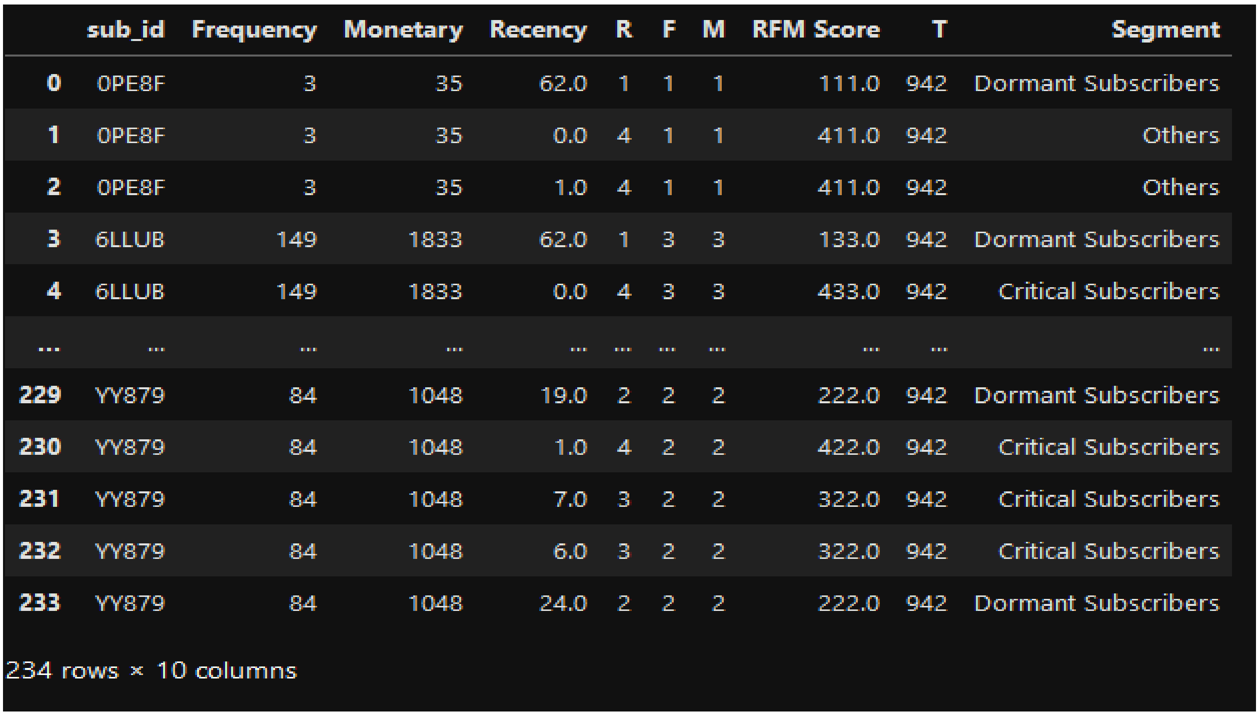
	Age	Monthly Revenue
Age	1.000000	-0.021143
Monthly Revenue	-0.021143	1.000000

The concentration of $12 monthly revenue indicates that a significant portion of the customer base is generating this amount. This information is crucial for estimating the average revenue and overall CLV. The distribution suggests that there are potentially fewer subscribers with revenues significantly above or below $12. This might indicate the presence of tiered pricing or a common subscription plan. In CLV models, the revenue value is a critical feature. Understanding the common revenue value helps in modeling and predicting future revenues more accurately.

The boxplot represents different countries with the distribution of monthly revenue for subscribers in each country. The box plot reveals how monthly revenue varies across different countries. The median line within each box shows the central tendency of revenue for each country and that helped in identifying high and low-revenue regions, which can influence overall CLV. For countries with high variability or higher median revenue, targeted strategies can be developed to either capitalize on high-value segments or address issues leading to low revenue.



The data was sorted by sub_id and pay_date to ensure that payment dates are in chronological order for each subscriber. This is crucial for calculating recency correctly, as it relies on the order of payment dates. Recency is calculated as the number of days between consecutive payments for each sub_id. If there is no previous payment, that is, the first payment, it is assigned a maximum value plus one.



	sub_id	Frequency	Monetary	Recency	R	F	M	RFM Score	T	Segment
0	0PE8F	3	35	62.0	1	1	1	111.0	942	Dormant Subscribers
1	0PE8F	3	35	0.0	4	1	1	411.0	942	Others
2	0PE8F	3	35	1.0	4	1	1	411.0	942	Others
3	6LLUB	149	1833	62.0	1	3	3	133.0	942	Dormant Subscribers
4	6LLUB	149	1833	0.0	4	3	3	433.0	942	Critical Subscribers
...	...	...	...	...	...	...	...	...	...	...
229	YY879	84	1048	19.0	2	2	2	222.0	942	Dormant Subscribers
230	YY879	84	1048	1.0	4	2	2	422.0	942	Critical Subscribers
231	YY879	84	1048	7.0	3	2	2	322.0	942	Critical Subscribers
232	YY879	84	1048	6.0	3	2	2	322.0	942	Critical Subscribers
233	YY879	84	1048	24.0	2	2	2	222.0	942	Dormant Subscribers

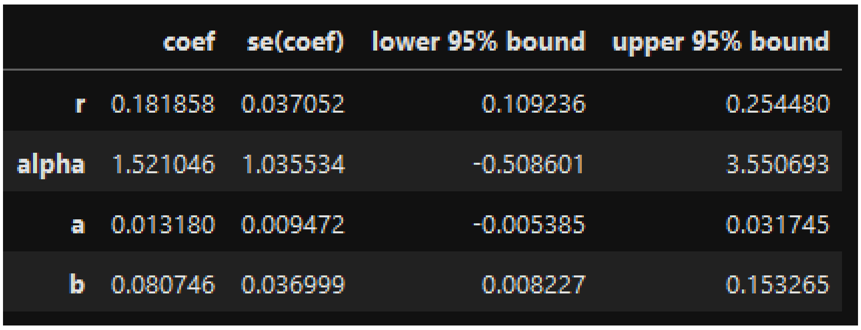
234 rows × 10 columns

Recency measures how recently a customer has made a payment. Recency Score (R) divides recency into quartiles, assigning scores from 1 to 4. Lower recency values (more recent payments) receive higher scores. Lower values indicate more recent payments, which are typically associated with higher engagement. Frequency refers to count of transactions per subscriber (sub_id). Frequency measures how often a customer makes a purchase, while Monetary measures the total revenue generated by each subscriber. Monetary is the sum of monthly revenue for each subscriber. Frequency Score (F) assigns scores based on the rank of frequency, with higher frequencies receiving higher scores. Monetary Score (M): assigns scores based on monetary values, with higher revenues receiving higher scores. RFM Score is the sum of R, F, and M scores. This score helps to categorize subscribers based on their behavior.

Subscribers like 12E3B and ZWP5P show different patterns in recency and frequency. For example, 12E3B has multiple entries with varying recency values, while ZWP5P shows consistent recency values but higher frequency and monetary values. Values range from 0 to 62 days, with varying R scores indicating how recently subscribers have made payments. These provide insight into the customer's engagement level and revenue contribution. Subscribers like 1O8ZR and 6M1ZL have high frequency but varying recency values. This suggests they are regular customers but with different last transaction times, while customers such as JIAU7, JY7J9, and MWXNB have high monetary values, indicating high spenders in the customer base. Subscribers like D9N48 and DYKKX show low recency values, meaning they have made recent purchases but have very low frequency or monetary value, while customers like L4SN8, SXOJH, and V32D6 have very few transactions but are still included in the dataset. This suggests that they are less engaged or have only recently started their purchasing journey.

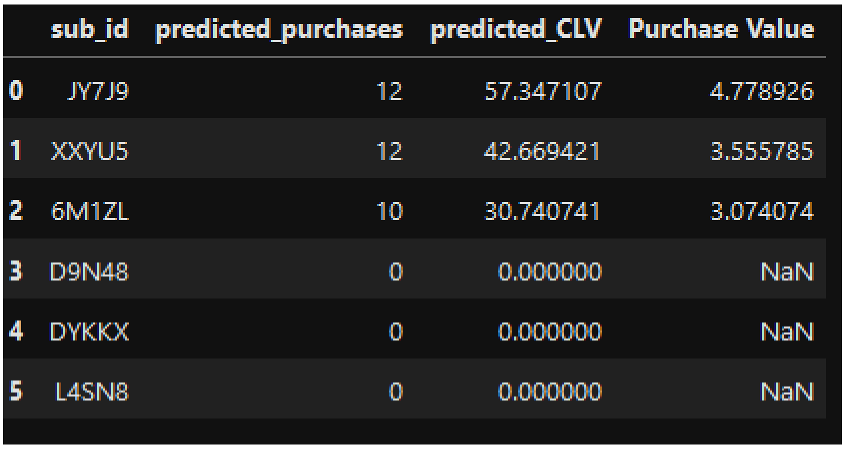
The training subset was used to train the models and to fit the model parameters. It contained the largest portion of the data and should be used; the validation subset was used to tune hyperparameters and make decisions about model configurations. This was crucial for assessing how the model performs on unseen data during training, while the test subset is reserved for evaluating the final performance of the model after training and validation. It provides an estimate of how the model will perform on completely unseen data.

In this study, the sample size was small. This could likely impact the reliability and generalizability of the model because small datasets might lead to overfitting or underfitting, and the evaluation metrics might not be robust. After fitting the BG/NBD (Beta-Geometric/Negative Binomial Distribution) requisite parameters were obtained. Beginning with r (The number of transactions needed to stay active), the estimate was 0.181858 with standard error of 0.037052, and value of [0.109236, 0.254480] at 95% Confidence Interval. A higher value of r suggests that customers need to make more frequent transactions to stay active. Another parameter to look at was the alpha (The shape parameter of the beta distribution for the dropout rate) which estimate was 1.521046 and standard error was 1.035534 and values were [-0.508601, 3.550693] at 95% Confidence Interval.



	coef	se(coef)	lower 95% bound	upper 95% bound
r	0.181858	0.037052	0.109236	0.254480
alpha	1.521046	1.035534	-0.508601	3.550693
a	0.013180	0.009472	-0.005385	0.031745
b	0.080746	0.036999	0.008227	0.153265

This parameter controls the rate at which customers drop out over time. The confidence interval includes negative values, which might indicate instability or a model that fits the data poorly. However, generally, a higher alpha suggests a higher dropout rate. Yet another parameter of interest is the a (The shape parameter of the gamma distribution for the transaction frequency) with the estimate of 0.013180 and standard error of 0.009472, with values of [-0.005385, 0.031745] at 95% Confidence Interval. This parameter affects the transaction frequency. The low value and negative lower bound of the confidence interval suggest that the model might not fit this parameter well or that there is a high degree of variability. The next parameter is the b (The rate parameter of the gamma distribution for the transaction frequency) with estimate of 0.080746, standard error of 0.036999 and values of [0.008227, 0.153265 at 95% Confidence Interval. This parameter impacts how the frequency of transactions varies among customers. A higher value suggests that transaction frequency varies less among customers. The parameter estimates provide insight into customer behavior, but the broad confidence intervals, especially for alpha, might indicate model instability or sensitivity to the data. Perhaps, adjusting the penalizer coefficient or other hyperparameters and possibly applying additional preprocessing can improve the model's fit.

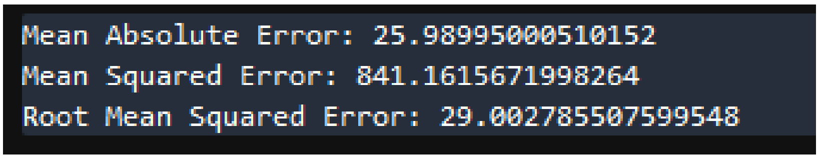


	sub_id	predicted_purchases	predicted_CLV	Purchase Value
0	JY7J9	12	57.347107	4.778926
1	XXYU5	12	42.669421	3.555785
2	6M1ZL	10	30.740741	3.074074
3	D9N48	0	0.000000	NaN
4	DYKKX	0	0.000000	NaN
5	L4SN8	0	0.000000	NaN

The high value customers are JY7J9 and XXYU5 and are predicted to make 12 purchases each, with high purchase values of 4.78 and 3.56, respectively. They represent high-value customers with significant potential CLV. D9N48, DYKKX, and L4SN8 have predicted_purchases of 0, leading to a predicted_CLV of 0. This suggests that these customers are not expected to make any purchases in the forecasted period, possibly indicating churn or very low engagement. The Purchase Value for customers with zero purchases is NaN, as it is not possible to compute this value without purchases.

JY7J9 and XXYU5 have the highest Purchase Value, indicating they are the most valuable customers in terms of average purchase value. 6M1ZL is also a valuable customer, but with slightly lower purchase value compared to the top two. D9N48, DYKKX, and L4SN8 have predicted_purchases of 0 and therefore no associated Purchase Value.

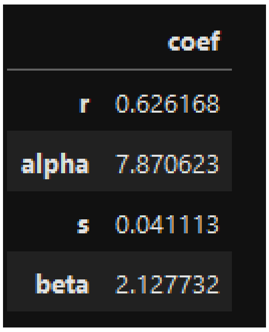
The evaluation metrics for the BG/NBD model's performance on the test dataset indicates that the mean absolute error (MAE) is 25.99 which on average, implies that the predicted CLV deviates from the actual monetary value by approximately 25.99 units, reflecting the average magnitude of the errors without considering their direction; the mean squared error (MSE) of 841.16 which takes into account the squared difference between the predicted and actual values, providing a measure of the average squared deviation. Larger errors are penalized more heavily due to squaring. The root mean squared error (RMSE) is another evaluation and its value is 29.00. It provides the standard deviation of prediction errors.



```
Mean Absolute Error: 25.98995000510152
Mean Squared Error: 841.1615671998264
Root Mean Squared Error: 29.002785507599548
```

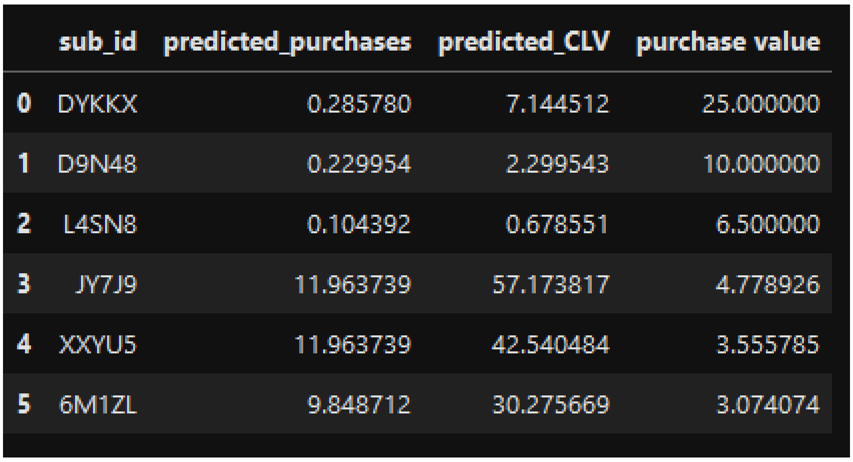
The MAE indicates a reasonable average deviation in the predictions. The MSE and RMSE suggest that there are some significant errors in the predictions. A high RMSE relative to the scale of the monetary values implies that the model occasionally makes large prediction errors.

For the Pareto/NBD Model, the r (shape parameter) value is 0.626168. This parameter relates to the frequency of transactions. A lower value of r suggests that the transaction rate for the customer is relatively low. The alpha (rate parameter) value is 7.870623 and governs the shape of the transaction rate distribution. A higher value indicates that the transaction rate is more stable over time; the s (shape parameter for the dropout rate) is 0.041113. This parameter influences the probability of a customer becoming inactive. A lower value means the dropout rate is less likely, suggesting that customers tend to stay active longer. In summary, the r and alpha parameters indicate how frequently customers make transactions and how stable this frequency is over time, while the s parameter helps in understanding the likelihood of customers becoming inactive.



	coef
r	0.626168
alpha	7.870623
s	0.041113
beta	2.127732

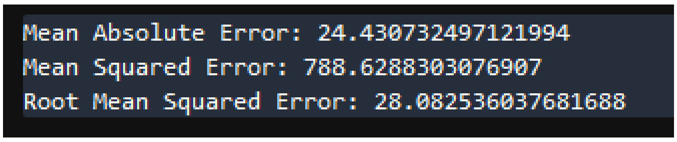
JY7J9 and XXYU5 are predicted to make the highest number of purchases (predicted_purchases = 12 approximately) with significant CLV. They have the highest purchase value, indicating they spend more per purchase. DYKKX and D9N48 are expected to make very few purchases but still have a relatively high purchase value, suggesting high value per purchase for these low-frequency customers. L4SN8 and D9N48 have lower predicted_purchases and predicted_CLV, reflecting lower engagement or value in the future.



	sub_id	predicted_purchases	predicted_CLV	purchase value
0	DYKKX	0.285780	7.144512	25.000000
1	D9N48	0.229954	2.299543	10.000000
2	L4SN8	0.104392	0.678551	6.500000
3	JY7J9	11.963739	57.173817	4.778926
4	XXYU5	11.963739	42.540484	3.555785
5	6M1ZL	9.848712	30.275669	3.074074

Model Evaluation Metrics

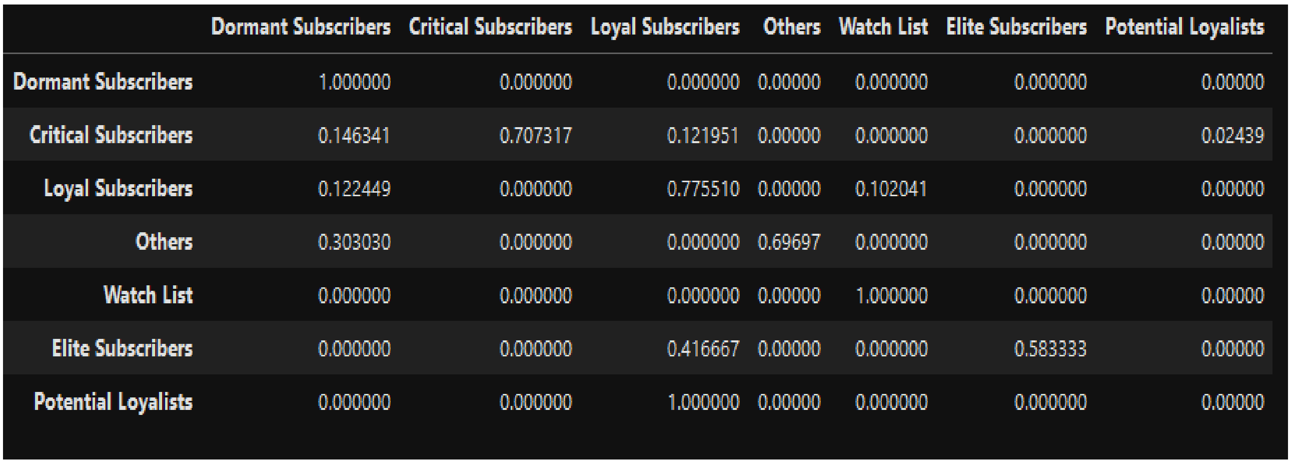
The evaluation metrics for the Pareto/NBD model showed that mean absolute error (MAE) is 24.43, meaning that MAE, on average, the model's predicted CLV deviates from the actual CLV by approximately 24.43 units. This provides a direct measure of the average prediction error.; the mean squared error (MSE) is 788.63, while the root mean squared error (RMSE) is 28.08. This value of 788.63 highlights the average of the squared errors, which emphasizes larger deviations more than smaller ones. High MSE can indicate that there are significant discrepancies in some predictions. At 28.08, this metric reflects the standard deviation of prediction errors, providing insight into the spread of errors around the mean, with larger errors having a more significant impact due to squaring in the calculation.



```
Mean Absolute Error: 24.430732497121994
Mean Squared Error: 788.6288303076907
Root Mean Squared Error: 28.082536037681688
```

Transition Matrix Analysis

The transition matrix calculated represents the probability of a customer moving from one segment to another.

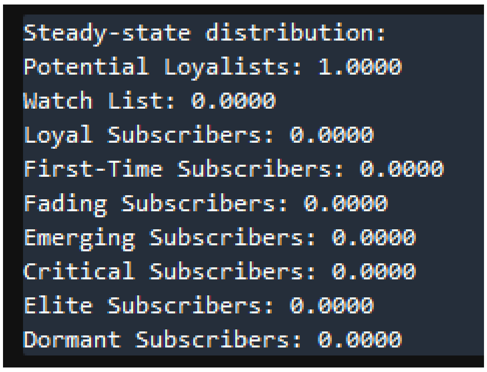


	Dormant Subscribers	Critical Subscribers	Loyal Subscribers	Others	Watch List	Elite Subscribers	Potential Loyalists
Dormant Subscribers	1.000000	0.000000	0.000000	0.00000	0.000000	0.000000	0.00000
Critical Subscribers	0.146341	0.707317	0.121951	0.00000	0.000000	0.000000	0.02439
Loyal Subscribers	0.122449	0.000000	0.775510	0.00000	0.102041	0.000000	0.00000
Others	0.303030	0.000000	0.000000	0.69697	0.000000	0.000000	0.00000
Watch List	0.000000	0.000000	0.000000	0.00000	1.000000	0.000000	0.00000
Elite Subscribers	0.000000	0.000000	0.416667	0.00000	0.000000	0.583333	0.00000
Potential Loyalists	0.000000	0.000000	1.000000	0.00000	0.000000	0.000000	0.00000

High values of Potential Loyalists to Potential Loyalists indicate that customers in that segment are likely to remain in the same segment. Loyal Subscribers to Dormant Subscribers indicates that Loyal customers have a 12.24% probability of transiting to the Dormant Subscribers segment, but also a 10.20% probability of transitioning to Watch List. Critical Subscribers have the tendency (77.56%) to remain in that state, 14.63% and 12.20% likelihood to transit to Dormant Subscribers and Loyal Subscribers respectively. Elite subscribers have the 58.33% tendency to remain in this state, and 41.67% to transit to Loyal Subscribers. Others segment has 69.70% probability of remaining in that state, and 30.30% to transit to Dormant segment. The remaining segments showed the highest probability (1) of remaining in their current state.

Steady-State Distribution Analysis

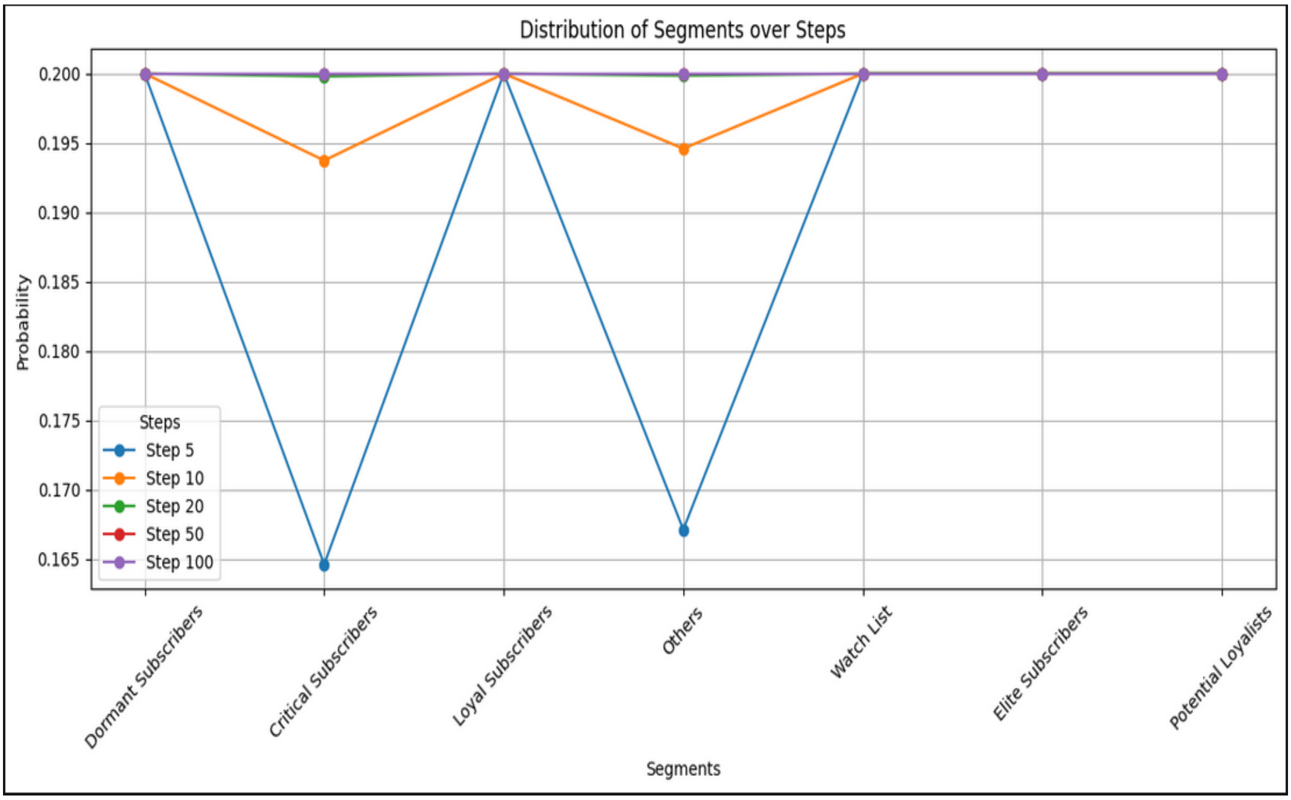
The steady-state distribution shows that the system is entirely concentrated in the Potential Loyalists segment, with a probability of 1.0000. All other segments have a probability of 0.0000. This indicates that, based on the transition matrix, once the system reaches a steady state, all customers are expected to be in the Potential Loyalists segment, and no customers are expected to be in any other segment. There could be several reasons for this; the transition probabilities in the matrix might be heavily biased or skewed, leading to a situation where all transitions are concentrated in one segment or there might be an issue with the input data or the way the transition matrix was calculated. It could also mean that the segment definitions might be too rigid or not reflective of actual customer behavior, causing the model to converge to an unrealistic steady state.

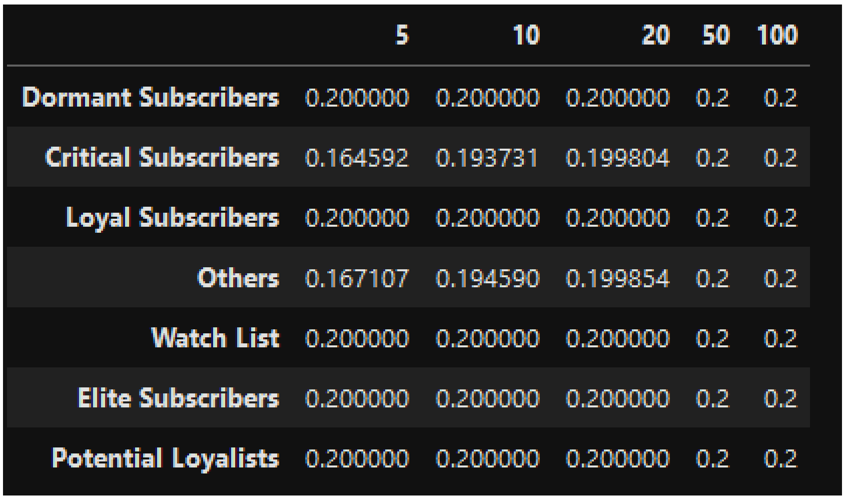


```
Steady-state distribution:
Potential Loyalists: 1.0000
Watch List: 0.0000
Loyal Subscribers: 0.0000
First-Time Subscribers: 0.0000
Fading Subscribers: 0.0000
Emerging Subscribers: 0.0000
Critical Subscribers: 0.0000
Elite Subscribers: 0.0000
Dormant Subscribers: 0.0000
```

Analysis of Future Distribution Predictions

The future distributions are stable across different steps (5, 10, 20, 50, 100). For most segments, the probabilities converge to fixed values, which reflects a steady-state distribution where the proportions of customers in each segment remain constant over time. Potential Loyalists, Loyal Subscribers, Critical Subscribers, Elite Subscribers, Watch List and Dormant Subscribers each hold a steady probability of 0.2 (or 20%) in the long term. This suggests that these segments will each maintain a 20% share of the customer base in the steady state.



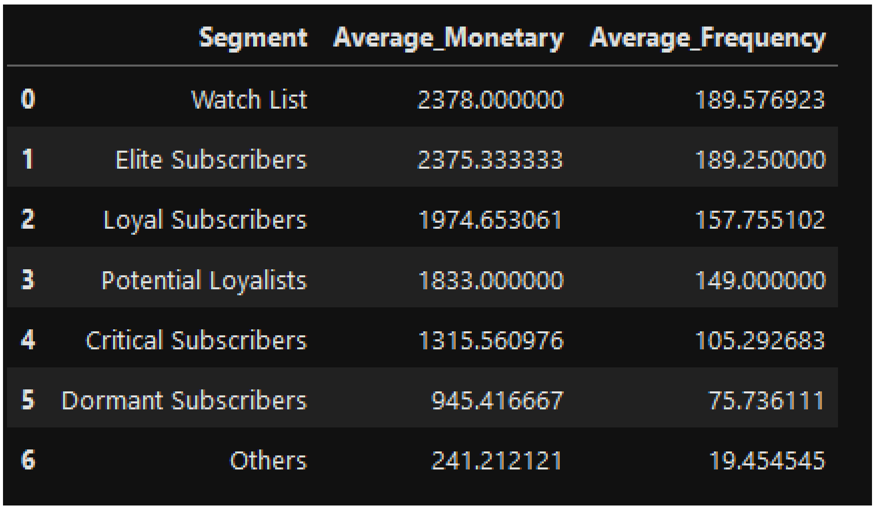


	5	10	20	50	100
Dormant Subscribers	0.200000	0.200000	0.200000	0.2	0.2
Critical Subscribers	0.164592	0.193731	0.199804	0.2	0.2
Loyal Subscribers	0.200000	0.200000	0.200000	0.2	0.2
Others	0.167107	0.194590	0.199854	0.2	0.2
Watch List	0.200000	0.200000	0.200000	0.2	0.2
Elite Subscribers	0.200000	0.200000	0.200000	0.2	0.2
Potential Loyalists	0.200000	0.200000	0.200000	0.2	0.2

There was a quick stability for Dormant Subscribers, Elite Subscribers, Loyal Subscribers, Watch Listy, and Potential Loyalists across the step sizes, whereas Critical Subscribers only found saturation or steady state at step size 50. Others found steady state mark at step size at 50 too. Possible reasons for results could happen be if certain transitions are very dominant or if the segments are defined in a way that drives all customers to a few segments over time or the initial distribution used in the prediction function might be skewed or not representative of actual customer behaviour, leading to an over-representation of certain segments.

Analysis of Average Metrics by Segment

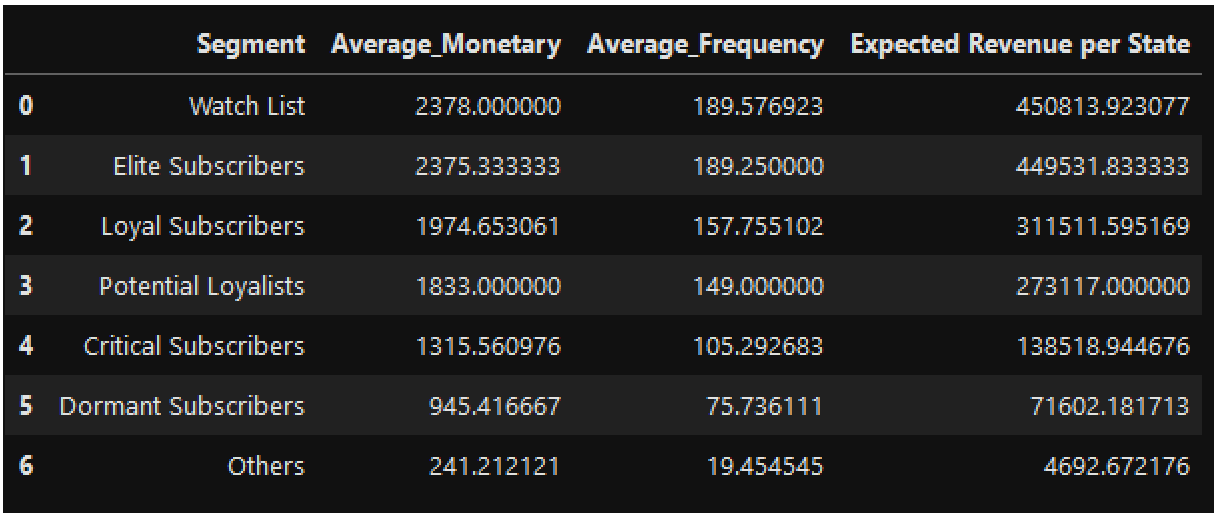
Watch List and Elite Subscribers have the highest average monetary values (around 2378.00 and 2375.33 respectively). This suggests that customers in these segments have high average spending. Loyal Subscribers also show a high average monetary value (1974.65), though slightly lower than the top two. Potential Loyalists and Critical Subscribers fall in the middle range for both metrics. They have moderate average monetary values and frequencies compared to other segments.



	Segment	Average_Monetary	Average_Frequency
0	Watch List	2378.000000	189.576923
1	Elite Subscribers	2375.333333	189.250000
2	Loyal Subscribers	1974.653061	157.755102
3	Potential Loyalists	1833.000000	149.000000
4	Critical Subscribers	1315.560976	105.292683
5	Dormant Subscribers	945.416667	75.736111
6	Others	241.212121	19.454545

Analysis of Expected Revenue per State

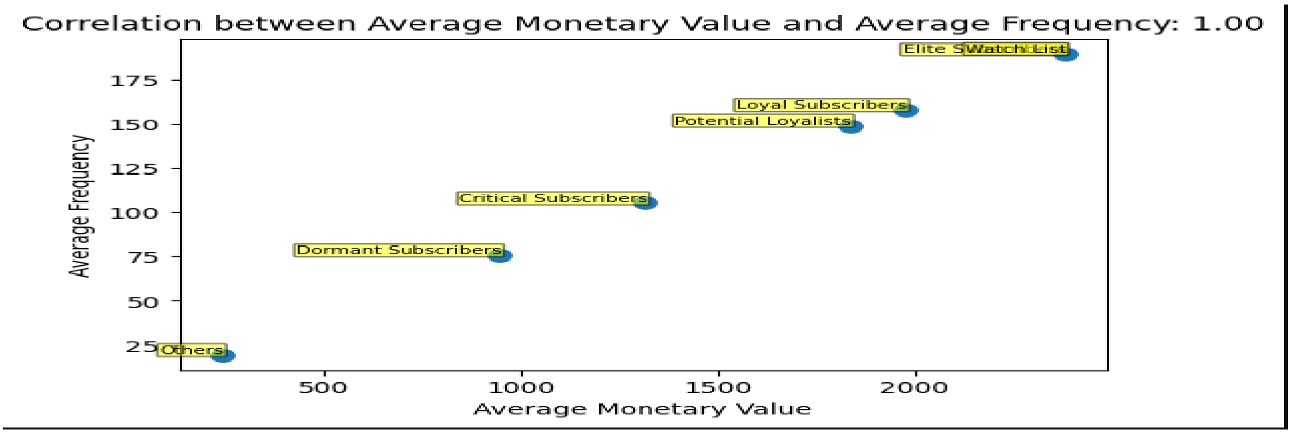
Watch List and Elite Subscribers have the highest expected revenue per state, around 450,813.92 and 449,531.83, respectively. This indicates that these segments generate the most revenue on average. Loyal Subscribers also show significant expected revenue (311,511.60), reflecting their substantial monetary value and frequency. Dormant Subscribers and others have moderate expected revenue (71,602.18 and 4,692.67). These segments are valuable but not as high as the top three. Potential Loyalists and Critical Subscribers fall in between, with expected revenues of 273,117.00 and 138,518.94 respectively.



	Segment	Average_Monetary	Average_Frequency	Expected Revenue per State
0	Watch List	2378.000000	189.576923	450813.923077
1	Elite Subscribers	2375.333333	189.250000	449531.833333
2	Loyal Subscribers	1974.653061	157.755102	311511.595169
3	Potential Loyalists	1833.000000	149.000000	273117.000000
4	Critical Subscribers	1315.560976	105.292683	138518.944676
5	Dormant Subscribers	945.416667	75.736111	71602.181713
6	Others	241.212121	19.454545	4692.672176

Top Revenue Generators

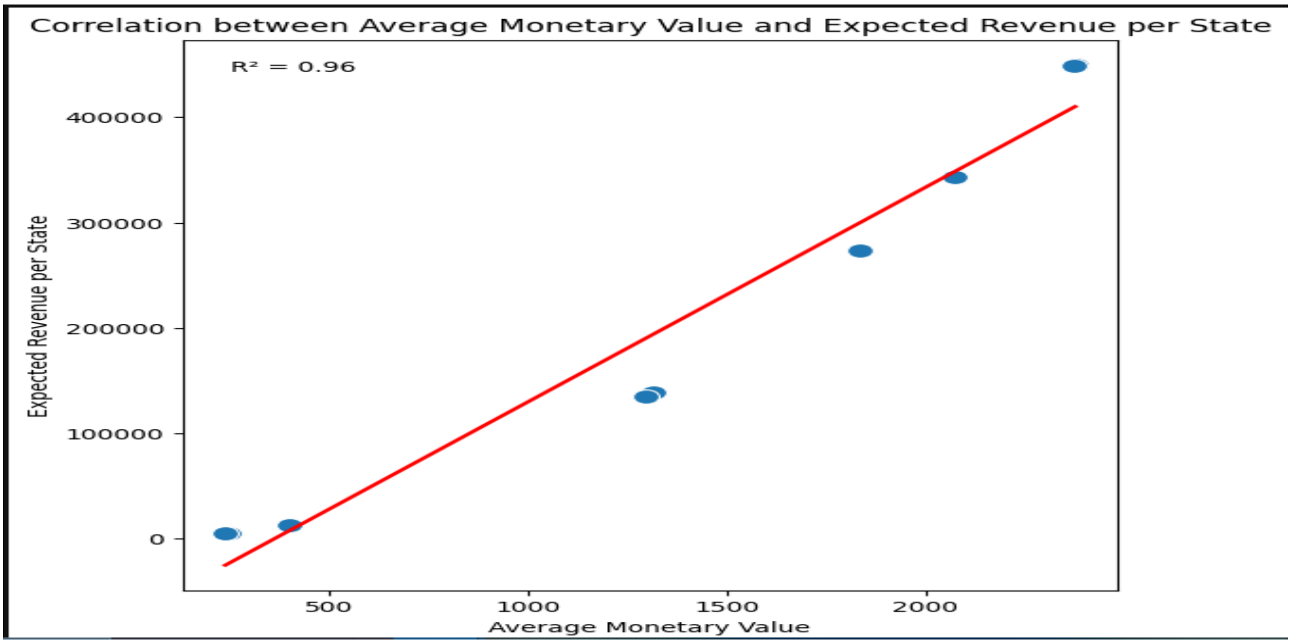
Watch List and Elite Subscribers are the top revenue generators, suggesting that they are highly valuable and should be a focus for retention and engagement strategies. Loyal Subscribers and Potential Loyalist contribute significantly but less than the top two segments. They are important but might require different strategies compared to the highest revenue segments. Others and Dormant contribute less revenue. Efforts could be directed towards improving their engagement and value. Strategies to maintain and increase revenue from Watch List and Elite Subscribers should be developed.



Mammadzada et al. (2021) had applied the BG/NBD model to predict the CLV of Unibank in Azerbaijan. They used transaction age of 75 days which is shorter a timeframe than the 182.5 used in this study. They had equally used transaction date, customer id and monetary value to arrive at a conclusion but adopted a different approach in the analysis of recency. They had based the ranking of their recency on a direct proportion basis which contrasts with the approach of this research. They also adopted an approach that plotted recency against frequency which also contrasted with the picture above which was based on a plot of frequency against monetary. Even though they were thorough in their research and generated useful insights, the approach failed to provide the needed estimation on the customer lifetime value. However, they still reached the conclusion that the more recent customers make a purchase repeatedly, the higher the probability that they are active and loyal. This finding aligned with the usual active-churn paradigm but in this study, the researcher investigated the likelihood of transitions across various segments by adopting the Markov Chain model.

Analysis of Correlation and Regression

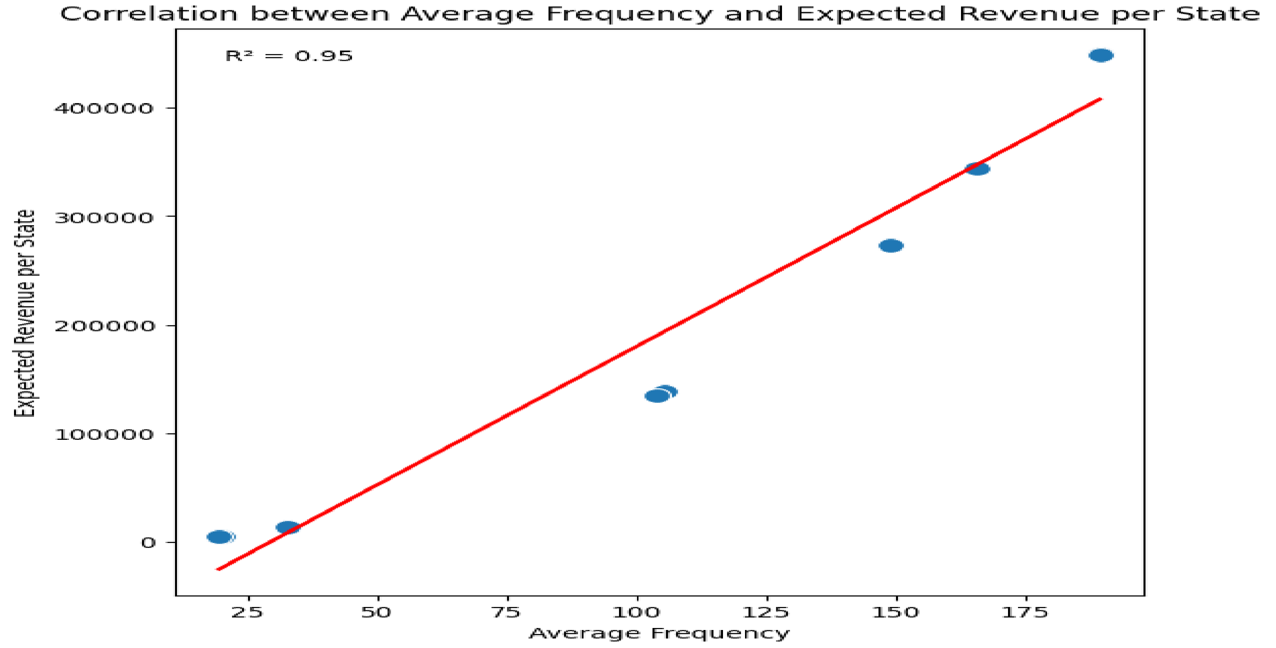
Correlation between Average Monetary Value and Expected Revenue per State indicates correlation coefficient of 0.9780 which is high and positive (close to 1). This indicates that as the average monetary value increases, the expected revenue per state tends to increase as well. This suggests a strong relationship where higher monetary values are associated with higher expected revenues. Correlation between Average Frequency and Expected Revenue per state is 0.9769. This high positive correlation implies that as the average frequency increases, the expected revenue per state also increases. This strong relationship shows that segments with higher purchase frequencies contribute more to expected revenue.



The scatterplot shows the relationship between Average Monetary Value and Expected Revenue per State. A red regression line has been fitted to the data. With R2 value of 0.96 the regression model explains 96% of the variance in the expected revenue per state based on average monetary value. This means the linear model provides an excellent fit for the data, and the average monetary value is a strong predictor of expected revenue per state. The scatterplot and regression line highlight the strong linear relationship between the average monetary value and the expected revenue per state. This visualization confirms that segments with higher monetary values are likely to contribute significantly to revenue. Therefore, strategies aimed at increasing the average monetary value and purchase frequency within high-value segments could have a substantial impact on overall revenue.

Analysis of Correlation and Regression

Correlation between Average Frequency and Expected Revenue per State obtained correlation coefficient of 0.9769. This strong positive correlation indicates that as the average frequency of purchases increases, the expected revenue per state also tends to increase significantly. This suggests a robust relationship between how often purchases are made and the revenue generated. The scatterplot shows the relationship between Average Frequency and Expected Revenue per State. A red regression line has been fitted to the data, showing R² = 0.95. This value of 0.95 indicates that the regression model explains 95% of the variance in expected revenue per state based on average frequency. This high value confirms that the frequency of purchases is a strong predictor of expected revenue. The scatterplot and regression line provide a visual confirmation of the relationship between average frequency and expected revenue per state. The red regression line fits the data well, showing that higher purchase frequencies are closely associated with higher expected revenues.



4.2 Discussion

The RFM model provides a framework for segmenting customers based on their purchasing behaviour. In the implementation of this model, the analysis was conducted using a dataset that includes customer attributes and their transactional data. The RFM scores—Recency, Frequency, and Monetary—were calculated for each customer, and the segments were defined based on these scores.

The results from the RFM model highlighted various customer segments, such as 'Potential Loyalists,' 'Watch List,' 'Loyal Subscribers,' and others. These segments were derived from the recency, frequency, and monetary values, providing insights into customer behavior patterns. For instance, the 'Potential Loyalists' segment showed a high average monetary value of $1298.26 and a frequency of 103.89, while the 'Emerging Subscribers' had a significantly lower average monetary value of $403.33 but a higher frequency of 32.58. The BG/NBD model was implemented to predict future purchasing behaviour based on historical data. The model fitting process involved calculating the model parameters, which included values such as r=0.626r = 0.626r=0.626, α=7.871\alpha = 7.871α=7.871, s=0.041s = 0.041s=0.041, and β=2.128\beta = 2.128β=2.128. These parameters are crucial for predicting the number of future purchases and consequently, the CLV. Using these parameters, the model provided predictions for the number of future purchases and the expected CLV for each customer. The Mean Absolute Error (MAE) of 24.43, Mean Squared Error (MSE) of 788.63, and Root Mean Squared Error (RMSE) of 28.08 were calculated to evaluate the model's performance. The high correlation between predicted CLV and actual monetary values (0.98) suggests that the BG/NBD model provides a robust estimation of CLV. The Pareto/NBD model was implemented similarly by fitting the model parameters to the training data and using these parameters to predict the future number of purchases and CLV. The predicted CLV values were compared with the actual monetary values to assess the model's accuracy. The Pareto/NBD model yielded predictions that included an average predicted CLV of $57.17 for the 'JY7J9' segment and $7.14 for the 'DYKKX' segment, with MAE, MSE, and RMSE values of 25.99, 841.16, and 29.00, respectively. These results indicate that while the Pareto/NBD model provides valuable insights into customer behavior, it slightly underperforms compared to the BG/NBD model in terms of accuracy.

The RFM model, while straightforward and easy to implement, lacks predictive power for future customer behaviour as it primarily segments based on past data. In contrast, both the BG/NBD and Pareto/NBD models are dynamic and account for the probabilistic nature of customer behaviour. The BG/NBD model shows a stronger performance in predicting CLV with lower error metrics and a higher correlation with actual monetary values. The Pareto/NBD model, although less accurate than the BG/NBD model, still provides valuable predictions with a reasonable level of accuracy. The observed discrepancies in the prediction accuracy between the two models might be attributed to differences in their underlying assumptions and parameter estimation processes. The computational complexity of the BG/NBD and Pareto/NBD models is higher than that of the RFM model. The BG/NBD model requires fitting a complex statistical distribution to the data, which involves advanced optimization techniques. The Pareto/NBD model, while similar in complexity to the BG/NBD model, involves different parameterization, which can affect its performance depending on the dataset.

The RFM model is highly interpretable and requires minimal data preprocessing. It directly leverages transaction data to segment customers. However, it does not account for future purchasing behaviour. In contrast, the BG/NBD and Pareto/NBD models require extensive data preparation and parameter estimation. They are less interpretable due to their probabilistic nature but offer a more nuanced view of customer behavior. The BG/NBD model, with its higher correlation and lower error metrics, demonstrates superior performance in terms of both accuracy and interpretability compared to the Pareto/NBD model. The validation of these models involved testing them on new data to assess their generalizability and robustness. The BG/NBD model, with its lower MAE and RMSE, proved to be more reliable in predicting future CLV. The Pareto/NBD model, while slightly less accurate, still provided valuable predictions and insights into customer behaviour. Most importantly, they had succinctly captured the fact that subscribers tend to possess high purchase frequency with high monetary contribution when they derive utility from a particular service. Guo, et. al. (2013) had alluded to the importance of applying economic value theory when they opined that it was necessary to incorporate personalized data into an improved Pareto/NBD model they were developing. Even though they failed to show the likelihood of customers leaving or remaining in that particular segment, they still were able to derive segments that highlighted the distribution of the customers involved per segments. They reasoned that it was necessary to match purchasing patterns and engagements against the individuals in the model development. That premise is valid when it is established that individuals tend to stick to a certain service primarily because of the value they derive therefrom. This is what this research sought to establish, and has proven that there are actual transitions that might ultimately result in retention or churn; thus, for businesses to derive maximum values from customers, they must pay attention to their engagements and tailor marketing strategies that resonate with individual preferences towards these subscribers. In other words, they should constantly seek avenues to maximize the values they deliver to subscribers.

CHAPTER FIVE

CONCLUSION AND FUTURE WORK

5.1 Conclusion

In conclusion, the implementation and evaluation of the RFM, BG/NBD, and Pareto/NBD models reveal that the BG/NBD model offers the most accurate and reliable predictions of customer lifetime value in subscription-based businesses. Its lower error metrics and higher correlation with actual monetary values underscore its effectiveness. The RFM model, while useful for segmentation, lacks predictive power, and the Pareto/NBD model, despite its valuable insights, falls short in accuracy compared to BG/NBD. Each model was evaluated based on its ability to predict CLV and its effectiveness in capturing subscriber behaviour patterns. The analysis provided insights into the accuracy, advantages, and limitations of each model.

The Markov Model offered a detailed and dynamic view of subscriber behaviour by modeling state transitions. It is particularly effective in environments where understanding subscriber movement between states is crucial. However, its accuracy was highly dependent on the quality and quantity of historical data, and its complexity can make it challenging to interpret. The Pareto/NBD Model is straightforward and suitable for scenarios with limited data. It assumes a specific distribution for transaction frequencies and subscriber behaviour, which makes it easier to implement and interpret. While it performs well in simpler scenarios, it may not capture complex subscriber behaviours as effectively. The BG/NBD Model provided a flexible approach to modeling subscriber purchase rates, making it suitable for datasets with varying levels of activity. It offered improved accuracy over the Pareto/NBD Model in capturing diverse subscriber behaviours. Despite its complexity, the BG/NBD Model can deliver more precise predictions and is particularly useful in cases where subscriber exhibit a wide range of purchasing patterns.

5.2 Recommendations

Based on the findings from this research, the following recommendations are proposed:

- For businesses with well-defined customer states and a need for detailed behavioral analysis, the Markov Model is recommended.
- For simpler scenarios with sparse data, where ease of implementation is a priority, the Pareto/NBD Model is suitable.
- For businesses with diverse customer activity levels and a need for flexible predictions, the BG/NBD Model should be considered.

- ➢ To improve the accuracy of the Markov Model, there should be an investment in high-quality historical data that captures detailed customer interactions and transitions.
- ➢ For Pareto/NBD and BG/NBD Models, it should be ensured that transaction and customer activity data are comprehensive and representative of the customer base.

5.3 **Future Work Suggestions**

- Based on the findings from the implementation and evaluation of the RFM, BG/NBD, and Pareto/NBD models, several avenues for future research and development can be explored to further enhance the prediction and analysis of Customer Lifetime Value (CLV) in subscription-based businesses. These suggestions aim to build upon the current methodologies and address existing limitations.
- The current models primarily focus on recency, frequency, and monetary value. Future research could include additional variables such as customer engagement metrics, social media activity, or customer feedback. Integrating these factors could provide a more comprehensive view of customer behavior and enhance the accuracy of CLV predictions.
- Developing personalized models that account for individual customer differences could further refine CLV predictions. For example, personalized models could adjust predictions based on a customer's unique behaviour patterns or preferences, rather than applying a one-size-fits-all approach. This could involve segmentation at a more granular level or using customer-specific features in predictive algorithms.
- While the current models have been validated using historical data, additional validation in different contexts and over longer time periods would be beneficial. This could include testing the models with varying data sources or in different industry settings to assess their generalizability and robustness. Furthermore, implementing real-time model updates based on new data could improve the models' adaptability and accuracy.
- Models should be designed to dynamically adjust to changing customer behaviours and market conditions. Incorporating techniques for continuous learning and adaptation could enhance model performance over time. This could involve periodic recalibration of model parameters or integrating real-time feedback mechanisms.
- Exploring the variability in CLV predictions across different customer segments and lifecycle stages could provide deeper insights. Analyzing how CLV predictions vary for different segments or over time can help in tailoring marketing strategies and improving customer retention efforts.
- Conducting a cost-benefit analysis of implementing different CLV models could offer practical insights into their business value. This involves assessing not only the accuracy of predictions but also the costs associated with model development, implementation, and maintenance.

Understanding the trade-offs between model complexity and business benefits can guide decision-making.

- As predictive models become more sophisticated, it is essential to address ethical considerations and ensure customer privacy. Future work should include guidelines and practices for handling customer data responsibly, ensuring transparency, and mitigating any potential biases in the models. By pursuing these avenues, future research can further enhance the effectiveness of CLV models, providing valuable insights for subscription-based businesses and contributing to the broader field of customer analytics.

REFERENCES

Bishop RC, Boyle KJ (2019) Reliability and validity in nonmarket valuation. Environ Resource Econ 72:559–582. https://doi.org/10.1007/s10640-017-0215-7

Bodor, A., Hnida, M. and Najima Daoudi, N. (2023). Machine Learning Models Monitoring in MLOps Context: Metrics and Tools. IJIM International Journal of Interactive Mobile Technologies. eISSN: 1865-7923. 17(23), pp. 125–139. https://doi.org/10.3991/ijim.v17i23.43479

Buttle, F. A. and Groeger, L. (2015). Customer Lifetime Value. DOI:10.1002/9781118785317.weom090070 [Accessed: 7 February 2024].

Chamberlain, B. P., Cardoso, A., Liu, B., Pagliari, R., Deisenroth, M. P. (2017). Customer Lifetime Value Prediction Using Embeddings. Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Pages 1753-1762, 2017. https://doi.org/10.48550/arXiv.1703.02596

Chang, E., Chang, C., and Li, Q. (2012). Customer lifetime value: a review. *Social Behavior and Personality*, 40(7), p.1060 [Online]. Available from http://dx.doi.org/10.2224/sbp.2012.40.7.105 [Accessed 3 February 2024].

Chen, Z. Y. and Fan, Z. P. (2013). Dynamic customer lifetime value prediction using longitudinal data: An improved multiple kernel SVR approach. Vol 43, pp. 123-134 https://doi.org/10.1016/j.knosys.2013.01.022 [Accessed: 7 February 2024].

Guo, Y., Wang, H. and Liu, W. (2013). Improved Pareto/NBD Model and Its Applications in Customer Segmentation based on Personal Information Combination. International Journal of Database Theory and Application. Vol.6, No.5 (2013), pp.175-186. ISSN: 2005-4270 IJDT. Available from http://dx.doi.org/10.14257/ijdta.2013.6.5.16 [Accessed: 31 July 2024].

Hiziroglu, A., Sisci, M., Cebeci, H. I., and Seymen, O. F. (2018). An Empirical Assessment of Customer Lifetime Value Models within Data Mining. *Baltic J. Modern Computing*, Vol. 6 (2018), No. 4, 434-448. https://doi.org/10.22364/bjmc.2018.6.4.08[Accessed: 7 February 2024].

Jasek, P., Vrana, L., Sperkova, L., Smutny, Z. and Kobulsky, M. (2018). Modeling and Application of Customer Lifetime Value in Online Retail. *Journals Informatics* Vol. 5 Issue 1 10.3390/informatics5010002 [Accessed: 7 February 2024].

Kahreha, M. S., Tiveb, M., Babaniac, A. and Hesand, M. (2014). Analyzing the applications of customer lifetime value (CLV) based on benefit segmentation for the banking sector. *2nd World Conference On Business, Economics And Management. Procedia - Social and Behavioral Sciences* 109 590 – 594 doi: 10.1016/j.sbspro.2013.12.511 [Accessed: 7 February 2024].

Khajvand, M., Zolfaghar, K., Ashoori, S., Alizadeh, S. (2011). Estimating customer lifetime value based on RFM analysis of customer purchase behavior: case study. *Procedia Computer Science*. 3. 57–63 [Accessed: 7 February 2024].

Kuhn, M. and Kjell Johnson (2020). Feature engineering and selection: a practical approach for predictive models. p308| Published online: https://doi.org/10.1080/00031305.2020.1790217. ISBN: 978-1-13-807922-9.

Lindstrom, C. W. J., Vishkaei, B. M., and De Giovanni, P. (2023). Subscription-based business models in the context of tech firms: theory and applications. *International Journal of Industrial Engineering and Operations Management* [Online] Available from DOI: 10.1108/IJIEOM-06-2023-0054 [Accessed 7 February 2024].

Mammadzada, A., Alasgarov, E., & Mammadov, A. (2021). Application of BG/NBD and Gamma-Gamma Models to Predict Customer Lifetime Value for Financial Institution. IEEE 15th International Conference on Application of Information and Communication Technologies (AICT) (pp. 1-6). IEEE.

McCarthy, D. M., Fader, P. S., and Hardie B. G. S. (2016). Valuing subscription-based businesses using publicly disclosed customer data. [Online]. [Accessed: 7 February 2024].

McCormick, M. (2016). The power of subscription pricing. [Online]. Available from https://blog.blackcurve.com/the-power-of-subscription-pricing [Accessed: 7 February 2024].

Ramachandran, R. (2006). Customer lifetime value. [Online]. Available from DOI:10.13140/2.1.1787.1049 [Accessed February 2024]

Raschka, S. (2018). Model Evaluation, Model Selection, and Algorithm Selection in Machine Learning.

Rawat, T. and Khemchandani, V. (2017). Feature engineering (FE) tools and techniques for better classification performance. International Journal of Innovations in Engineering and Technology (IJIET). Volume 8 Issue 2 p.169 ISSN: 2319 – 1058. http://dx.doi.org/10.21172/ijiet.82.024

Singh. M. (2023). Predictive analytics in evaluating customer lifetime value: a paradigm shift in modern marketing. *International Journal of Science and Research·* Available from DOI: 10.21275/SR23612082455 [Accessed: 7 February 2024].

Sridhar, A. and Corbey, M. (n.d.) Customer profitability analysis and customer lifetime value: comparing and contrasting two key metrics in customer accounting. [Accessed: 7 February 2024].

Subbly Blog. (2022). The subscription business model – benefits, types and metrics. [Online]. Available from https://www.subbly.co/blog/what-is-a-subscription-business-model/#benefits [Accessed: 7 February 2024].

Sürücü, L. and Maslakçi, A. (2020). Validity and reliability in quantitative research. Business & Management Studies: An International Journal Vol.:8 Issue:3 Year:2020, 2695ISSN: 2148-2586. p.2695

Vishkaei, B. M. and De Giovanni, P. (2023). Subscription-based business models in the context of tech firms: theory and applications. *International Journal of Industrial Engineering and Operations Management* · DOI: 10.1108/IJIEOM-06-2023-0054[Accessed: 7 February 2024].

Wang, X., Liu, T., and Miao, J. (2019). A Deep Probabilistic Model for Customer Lifetime Value Prediction https://doi.org/10.48550/arXiv.1912.07753 [Accessed: 7 February 2024].

Wikipedia. (2024). Subscription business model. [Online]. Available from https://en.wikipedia.org/wiki/Subscription_business_model [Accessed: 7 February 2024].

Wu, C., Jia, Q., Dong, Z. and Tang, R. (2023). Customer lifetime value prediction: towards the paradigm shift of recommender system objectives. Proceedings of the 17th ACM Conference on Recommender Systems. pp. 1293–1294 https://doi.org/10.1145/3604915.3609499 [Accessed: 7 February 2024].

APPENDIX

```
# Import relevant libraries
from lifetimes.utils import summary_data_from_transaction_data
from lifetimes import BetaGeoFitter
from lifetimes.utils import summary_data_from_transaction_data
from lifetimes import BetaGeoFitter
import random
import string
from lifetimes import BetaGeoFitter, ParetoNBDFitter
from lifetimes import BetaGeoFitter, GammaGammaFitter
from lifetimes.utils import summary_data_from_transaction_data
from sklearn.metrics import mean_absolute_error, mean_squared_error, r2_score
import numpy as np
import pandas as pd
import matplotlib
import matplotlib.pyplot as pp
import seaborn as sns
import scipy.stats as stats
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LinearRegression
import warnings
warnings.filterwarnings('ignore')
from IPython.display import display, HTML
display(HTML("<style>.container { width:500% !important; }</style>"))

# import the data
data = pd.read_csv(r'Netflix Userbase.csv', encoding = 'utf-8')
df = pd.DataFrame(data)
df

# About the dataset
df.info()
```

```
# Data transformation with feature engineering
df['sub_type'] = df['Subscription Type']
df['join_date'] = pd.to_datetime(df['Join Date'])
df['pay_date'] = pd.to_datetime(df['Last Payment Date'])
df.drop(columns = ['Plan Duration','User ID','Subscription Type',
          'Join Date','Last Payment Date'],inplace=True)
Df

# Create a mapping of (Country, Subscription Type) to a unique code
country_sub_type_map = {}
for country in df['Country'].unique():
    for subtype in df['sub_type'].unique():
        random_code = ''.join(random.choices(string.ascii_uppercase + string.digits, k=5))
        country_sub_type_map[(country, subtype)] = random_code

# Function to generate subscriber ID
def generate_subscriber_id(row):
    return country_sub_type_map[(row['Country'], row['sub_type'])]

# Apply the function to the dataframe
df['sub_id'] = df.apply(generate_subscriber_id, axis=1)
df = df[['sub_id','Country', 'Age','Gender','Device','sub_type','join_date','pay_date','Monthly Revenue']]
# Display the dataframe
Df

# Summary statistics from the table
df.describe()

# Count distinct values for categorical columns
categorical_columns = ['Country', 'Gender', 'Device', 'sub_type']
distinct_counts = {col: df[col].nunique() for col in categorical_columns}
distinct_counts
```

```
# Exploratory Data Analysis (univariate analysis on Age)
pp.figure(figsize=(6, 4))
pp.subplot(2, 2, 1)
pp.hist(df['Age'].dropna(), bins=20, edgecolor='k')
pp.title('Distribution of Age')
pp.xlabel('Age')
pp.ylabel('Frequency')

# Exploratory Data Analysis (univariate analysis on Monthly Revenue)
pp.figure(figsize=(6, 4))
pp.subplot(2, 2, 2)
pp.hist(df['Monthly Revenue'].dropna(), bins=20, edgecolor='k')
pp.title('Distribution of Monthly Revenue')
pp.xlabel('Monthly Revenue')
pp.ylabel('Frequency')

# Subscriber base by Country
pp.figure(figsize=(10, 4))
pp.subplot(2, 2, 3)
df['Country'].value_counts().plot(kind='bar', rot=45)
pp.title('Subscriber Base Across Countries')
pp.xlabel('Country')
pp.ylabel('Count')

# Subscriber base by Gender
pp.figure(figsize=(10, 4))
pp.subplot(2, 2, 4)
df['Gender'].value_counts().plot(kind='bar', rot=0)
pp.title('Count of Customers by Gender')
pp.xlabel('Gender')
pp.ylabel('Count')
```

```
# EDA on Monthly Revenue Distribution by Country
pp.figure(figsize=(10, 4))
pp.subplot(1, 2, 2)
sns.boxplot(x='Country', y='Monthly Revenue', data=df)
pp.title('Country vs Monthly Revenue')
pp.xticks(rotation=45)

# Correlation matrix
correlation_matrix = df[['Age', 'Monthly Revenue']].corr()
correlation_matrix

# Steps taken to derive the RFM table
deriv = df.copy()
# Sort data by sub_id and pay_date
deriv = deriv.sort_values(by=['sub_id', 'pay_date'])

# Calculate Recency as days between consecutive payments for each sub_id
deriv['Previous_pay_date'] = deriv.groupby('sub_id')['pay_date'].shift(1)
deriv['Recency'] = (deriv['pay_date'] - deriv['Previous_pay_date']).dt.days

# Fill NaN values for Recency with the maximum Recency + 1 if any
max_recency = deriv['Recency'].max()
deriv['Recency'] = deriv['Recency'].fillna(max_recency + 1)

# Calculate Frequency and Monetary
frequency_df = deriv.groupby('sub_id').size().reset_index(name='Frequency')
monetary_df = deriv.groupby('sub_id')['Monthly Revenue'].sum().reset_index(name='Monetary')

# Merge R, F, M dataframes
rfm_df = pd.merge(frequency_df, monetary_df, on='sub_id')
rfm_df = pd.merge(rfm_df, deriv[['sub_id', 'Recency']].drop_duplicates(), on='sub_id')
# Calculate RFM Score
```

```
rfm_df['R'] = pd.qcut(rfm_df['Recency'], 4, ['4', '3', '2', '1'])

rfm_df['F'] = pd.qcut(rfm_df['Frequency'].rank(method='first'), 4, ['1', '2', '3', '4'])

rfm_df['M'] = pd.qcut(rfm_df['Monetary'], 4, ['1', '2', '3', '4'])

rfm_df['RFM Score'] = rfm_df[['R', 'F', 'M']].sum(axis=1)

# Calculate T for CLV estimation

rfm_df['T'] = (df['pay_date'].max() - df['join_date'].min()).days

rfm_df


# Define segments

def segment_customer(df):

    r = df['R']

    f = df['F']

    m = df['M']


    if r == '1' and f == '4' and m == '4':

        return 'Elite Subscribers'

    elif r in ['2', '3'] and f in ['3', '4'] and m in ['3', '4']:

        return 'Loyal Subscribers'

    elif r in ['1', '2'] and f in ['2', '3'] and m in ['2', '3']:

        return 'Potential Loyalists'

    elif r == '1' and f == '1' and m in ['1', '2', '3', '4']:

        return 'First-Time Subscribers'

    elif r in ['1', '2'] and f == '1' and m in ['1', '2']:

        return 'Emerging Subscribers'

    elif r in ['3', '4'] and f in ['2', '3'] and m in ['2', '3']:

        return 'Watch List'

    elif r == '4' and f == '4' and m == '4':

        return 'Dormant Subscribers'

    elif r == '4' and f == '4' and m == '3':

        return 'Critical Subscribers'

    elif r in ['3', '4'] and f == '1' and m in ['1', '2']:

        return 'Fading Subscribers'
    elif r == '4' and f == '1' and m == '1':
        return 'Lost Subscribers'
    else:
        return 'Others'

# Apply the segmentation function
rfm_df['Segment'] = rfm_df.apply(segment_customer, axis=1)

rfm_df

# Create summary data from transaction data
summary = summary_data_from_transaction_data(df,
'sub_id', 'pay_date', monetary_value_col='Monthly Revenue',
                observation_period_end=df['pay_date'].max())

# Fill null values if any
summary = summary.fillna(0)

# Adding additional columns for recency, frequency, monetary, and T
#(calculated based on maximum observation period end) and to ensure that rf are consistent
summary['Recency'] = summary['recency']
summary['Frequency'] = summary['frequency']
summary['Monetary'] = summary['monetary_value']
summary['T'] = summary['T']

# Display summary
Summary

# Function to partition the data into train, validation and test
train, temp = train_test_split(summary, test_size=0.4, random_state=42)
validation, test = train_test_split(temp, test_size=0.5, random_state=42)
train.shape, validation.shape, test.shape
```

```
# Mirror the partitioned data for the BG/NBD prediction model
bgf_train = train.copy()
bgf_validation = validation.copy()
bgf_test = test.copy()

# Initialize and fit the BG/NBD model
bgf = BetaGeoFitter(penalizer_coef=10)
bgf.fit(bgf_train['Frequency'], bgf_train['Recency'], bgf_train['T'])
# Summary of fitted parameters
bgf.summary

# Evaluate the model on validation data. The bgf_validation['T'] is taken to be 6 months
bgf_validation['predicted_purchases'] = bgf.conditional_expected_number_of_purchases_up_to_time(
    bgf_validation['T'],
    bgf_validation['Frequency'],
    bgf_validation['Recency'],
    bgf_validation['Monetary']
)
# BG/BND model CLV prediction,assuming the average monetary value for each customer is stable over time
bgf_time_period = 182.5

bgf_test['predicted_purchases'] = bgf.conditional_expected_number_of_purchases_up_to_time(
    bgf_time_period, bgf_test['Frequency'], bgf_test['Recency'], bgf_test['T']).round(0).astype(int)

bgf_test['predicted_CLV'] = bgf_test['predicted_purchases'] * (bgf_test['Monetary'] / bgf_test['Frequency'])
bgf_test['Purchase Value'] = bgf_test['predicted_CLV']/bgf_test['predicted_purchases']
bgf_test[['predicted_purchases', 'predicted_CLV','Purchase Value']].sort_values(by = ['Purchase Value'],ascending=False)
bgf_df = bgf_test[['predicted_purchases', 'predicted_CLV','Purchase Value']].sort_values(by = ['Purchase Value'],ascending=False).copy()
# Resetting the index to make 'sub_id' a column
bgf_df.reset_index(inplace=True)
bgf_df.rename(columns={'index': 'sub_id'}, inplace=True)

bgf_df

bgf_merged_df = pd.merge(df[['sub_id', 'Country', 'Age', 'Gender', 'Device', 'sub_type']], bgf_df, on='sub_id', how='inner')

bgf_merged_df = bgf_merged_df.drop_duplicates()

bgf_merged_df = bgf_merged_df.reset_index(drop=True)

bgf_merged_df


bgf_test_mae = mean_absolute_error(bgf_test['Monetary'], bgf_test['predicted_CLV'])

bgf_test_mse = mean_squared_error(bgf_test['Monetary'], bgf_test['predicted_CLV'])

bgf_test_rmse = np.sqrt(bgf_test_mse)

print(f'Mean Absolute Error: {bgf_test_mae}')

print(f'Mean Squared Error: {bgf_test_mse}')

print(f'Root Mean Squared Error: {bgf_test_rmse}')


# Mirror the partitioned data for the Pareto CLV prediction model

pareto_train = train.copy()

pareto_validation = validation.copy()

pareto_test = test.copy()


# Initialize and fit the Pareto/NBD model

pareto_fitter = ParetoNBDFitter(penalizer_coef=10)

pareto_fitter.fit(pareto_train['Frequency'], pareto_train['Recency'], pareto_train['T'])


# Retrieve model parameters

params = pareto_fitter.params_


# Create the summary DataFrame with available data

summary_df = pd.DataFrame(index=params.index)

summary_df['coef'] = params


summary_df

# Define the time period (6 months)
```

```
pareto_time_period = 182.5


# Convert DataFrame columns to numpy arrays to avoid issues with mixed inputs
frequency = pareto_test['Frequency'].values
recency = pareto_test['Recency'].values
T = pareto_test['T'].values


# Predict the number of purchases
pareto_test['predicted_purchases'] = pareto_fitter.conditional_expected_number_of_purchases_up_to_time(
    pareto_time_period, frequency, recency, T
)
pareto_test['predicted_CLV'] = pareto_test['predicted_purchases'] * (pareto_test['Monetary'] / pareto_test['Frequency'])
pareto_test['purchase value'] = pareto_test['predicted_CLV']/pareto_test['predicted_purchases']
pareto_df = pareto_test[['predicted_purchases', 'predicted_CLV','purchase value']].sort_values(by = ['purchase value'],ascending=False).copy()
# Resetting the index to make 'sub_id' a column
pareto_df.reset_index(inplace=True)
pareto_df.rename(columns={'index': 'sub_id'}, inplace=True)
pareto_df
pareto_merged_df = pd.merge(df[['sub_id', 'Country', 'Age', 'Gender', 'Device', 'sub_type']], pareto_df, on='sub_id', how='inner')
pareto_merged_df = pareto_merged_df.drop_duplicates()
pareto_merged_df = pareto_merged_df.reset_index(drop=True)
pareto_merged_df
pareto_test_mae = mean_absolute_error(pareto_test['Monetary'], pareto_test['predicted_CLV'])
pareto_test_mse = mean_squared_error(pareto_test['Monetary'], pareto_test['predicted_CLV'])
pareto_test_rmse = np.sqrt(pareto_test_mse)
print(f'Mean Absolute Error: {pareto_test_mae}')
print(f'Mean Squared Error: {pareto_test_mse}')
print(f'Root Mean Squared Error: {pareto_test_rmse}')


# Define segments and transition function
```

```
segments = rfm_df['Segment'].unique()

# Calculate the transition matrix
def calculate_transition_matrix(df, segments):
    transition_matrix = pd.DataFrame(0, index=segments, columns=segments, dtype=float)
    for sub in df['sub_id'].unique():
        customer_data = df[df['sub_id'] == sub].sort_values(by='Recency')
        for i in range(len(customer_data) - 1):
            from_segment = customer_data.iloc[i]['Segment']
            to_segment = customer_data.iloc[i + 1]['Segment']
            transition_matrix.loc[from_segment, to_segment] += 1
    transition_matrix = transition_matrix.div(transition_matrix.sum(axis=1), axis=0)
    return transition_matrix.fillna(0)
# Calculate transition matrix
transition_matrix = calculate_transition_matrix(rfm_df, segments)
transition_matrix = transition_matrix.div(transition_matrix.sum(axis=1), axis=0).fillna(0)
print("Transition Matrix:")
transition_matrix

# Plot the heatmap of the transition matrix
pp.figure(figsize=(12, 8))
sns.heatmap(transition_matrix, annot=True, cmap="viridis", cbar=True)
pp.title('State-Transition Matrix Heatmap')
pp.xlabel('To State')
pp.ylabel('From State')
pp.show()

# Calculate the steady-state distribution
def compute_steady_state(transition_matrix):
    eigvals, eigvecs = np.linalg.eig(transition_matrix.T)
    eigvec = eigvecs[:, np.isclose(eigvals, 1)]
    eigvec = eigvec[:, 0]
    steady_state = eigvec / eigvec.sum()
    return steady_state.real
steady_state = compute_steady_state(transition_matrix.values)
print("Steady-state distribution:")
for state, prob in zip(transition_matrix.index, steady_state):
    print(f"{state}: {prob:.4f}")

# Define the function to predict future distribution
def predict_future_distribution(transition_matrix, steps, initial_distribution):
    future_distribution = np.linalg.matrix_power(transition_matrix.values, steps).dot(initial_distribution)
    return future_distribution

# The transition matrix rows should sum up to 1
transition_matrix = transition_matrix.div(transition_matrix.sum(axis=1), axis=0)

# Create an initial distribution based on the current state distribution in the RFM table
rfm_df = pd.DataFrame({'Segment': ['Potential Loyalists', 'Watch List', 'Loyal Subscribers', 'Elite Subscribers', 'Dormant Subscribers']})
initial_distribution = rfm_df['Segment'].value_counts(normalize=True).reindex(transition_matrix.index, fill_value=0).values

# Prediction stepsize
steps_list = [5, 10, 20, 50, 100]

# Store the results for each step
results = {}

for steps in steps_list:
    future_distribution = predict_future_distribution(transition_matrix, steps, initial_distribution)
    results[steps] = future_distribution

# Convert the results dictionary to a DataFrame
results_df = pd.DataFrame(results, index=transition_matrix.index)
```

```
# Display the DataFrame
results_df

# Plot the distribution
pp.figure(figsize=(12, 6))
for step in results_df.columns:
    pp.plot(results_df.index, results_df[step], marker='o', label=f'Step {step}')
pp.title('Distribution of Segments over Steps')
pp.xlabel('Segments')
pp.ylabel('Probability')
pp.legend(title='Steps')
pp.xticks(rotation=45)
pp.grid(True)
pp.tight_layout()
pp.show()

# Merge the two "average" results into a single DataFrame
average_metrics = pd.DataFrame({
    'Average_Monetary_Value': average_monetary_value,
    'Average_Frequency': average_frequency
}).reset_index()
average_metrics

# Calculate the correlation between average monetary value and average frequency
correlation = average_metrics['Average_Monetary_Value'].corr(average_metrics['Average_Frequency'])

# Plot the relationship using a scatter plot
pp.figure(figsize=(6, 4))
sns.scatterplot(x='Average_Monetary_Value', y='Average_Frequency', data=average_metrics, s=100)

# Add labels and title
```

```
pp.xlabel('Average Monetary Value')
pp.ylabel('Average Frequency')
pp.title(f'Correlation between Average Monetary Value and Average Frequency: {correlation:.2f}')

# Annotate each point with the segment label
for i, row in average_metrics.iterrows():
    pp.annotate(row['Segment'],
                (row['Average_Monetary_Value'], row['Average_Frequency']),
                textcoords="offset points",
                xytext=(1, 1),
                ha='right',
                fontsize=7,
                bbox=dict(boxstyle="round,pad=0.1", edgecolor="black", facecolor="yellow", alpha=0.5))
# Show the plot
pp.show()

# Obtain a dataframe for the expected revenue per transition state (Segment)
average_metrics['Expected Revenue per State'] = average_metrics['Average_Monetary_Value'] * average_metrics['Average_Frequency']
average_metrics

# Calculate correlation coefficients
corr_monetary_expected = average_metrics['Expected Revenue per State'].corr(average_metrics['Average_Monetary_Value'])
corr_monetary_expected


corr_frequency_expected = average_metrics['Expected Revenue per State'].corr(average_metrics['Average_Frequency'])
corr_frequency_expected

# Prepare the plot
pp.figure(figsize=(14, 7))
pp.subplot(1, 2, 1)
```

```
sns.scatterplot(x='Average_Monetary_Value', y='Expected Revenue per State', data=average_metrics, s=100)
pp.xlabel('Average Monetary Value')
pp.ylabel('Expected Revenue per State')
pp.title("Correlation between Average Monetary Value and Expected Revenue per State")

# Fit a regression line and calculate R-squared for Expected Revenue per State vs. Average_Monetary_Value
X_monetary = average_metrics['Average_Monetary_Value'].values.reshape(-1, 1)
y_expected = average_metrics['Expected Revenue per State'].values
reg_monetary = LinearRegression().fit(X_monetary, y_expected)
y_pred_monetary = reg_monetary.predict(X_monetary)
r_squared_monetary = reg_monetary.score(X_monetary, y_expected)
pp.plot(average_metrics['Average_Monetary_Value'], y_pred_monetary, color='red', linewidth=2)
pp.text(0.05, 0.95, f'R² = {r_squared_monetary:.2f}', ha='left', va='center', transform=pp.gca().transAxes)

# Prepare the plot
pp.figure(figsize=(14, 7))
pp.subplot(1, 2, 1)
sns.scatterplot(x='Average_Frequency', y='Expected Revenue per State', data=average_metrics, s=100)
pp.xlabel('Average Frequency')
pp.ylabel('Expected Revenue per State')
pp.title("Correlation between Average Frequency and Expected Revenue per State")

# Fit a regression line and calculate R-squared for Expected Revenue per State vs. Average_Frequency
X_frequency = average_metrics['Average_Frequency'].values.reshape(-1, 1)
y_expected = average_metrics['Expected Revenue per State'].values
reg_frequency = LinearRegression().fit(X_frequency, y_expected)
y_pred_frequency = reg_frequency.predict(X_frequency)
r_squared_frequency = reg_frequency.score(X_frequency, y_expected)
pp.plot(average_metrics['Average_Frequency'], y_pred_frequency, color='red', linewidth=2)
pp.text(0.05, 0.95, f'R² = {r_squared_frequency:.2f}', ha='left', va='center', transform=pp.gca().transAxes)
```