

MRP

by Aarnav Aggarwal

Submission date: 03-May-2026 07:44PM (UTC+0530)

Submission ID: 2950548991

File name: Arnav_Aggarwal_24_BMBA_05.pdf (739.15K)

Word count: 13946

Character count: 84207

Project Dissertation Report On

Customer Lifetime Value Prediction for Retail Analytics: A Comparative Study of Machine Learning Models

Submitted By :

Arnav Aggarwal

24/BMBA/05

Under the Guidance of :

Dr. Yogesh Sharma

Assistant Professor



DELHI SCHOOL OF MANAGEMENT

Delhi Technological University

Bawana Road, Delhi - 110042

CERTIFICATE

This is to certify that the project titled “Customer Lifetime Value Prediction for retail analytics : A comparative study of machine learning models” is a bonafide work carried out by Amav Aggarwal under my guidance in partial fulfillment of the requirements for the award of the degree of MBA at Delhi Technological University.

Date:

Place:

65

(Signature of Guide)

Name: Dr. Yogesh Sharma

Designation: Assistant Professor

DECLARATION

I² hereby declare that the project report titled “Customer Lifetime Value Prediction for retail analytics : A comparative study of machine learning models”⁹ is my original work and has been carried out under the guidance of my faculty mentor.

Date:

Place:

(Signature)

Arnav Aggarwal

24/BMBA/05

I would like to express my sincere gratitude to my faculty guide for their valuable guidance, continuous support, and encouragement throughout the completion of this project.

I am also thankful to Delhi School of Management, Delhi Technological University, for providing the necessary resources and academic environment to carry out this research work.

Lastly, I would like to thank my family and friends for their constant support and motivation.

Arnav Aggarwal

24/BMBA/05

EXECUTIVE SUMMARY

Modern retail organizations are confronted with an increasingly complex question at the heart of their marketing strategy: which customers deserve the greatest investment of time, money, and attention? The answer, in its most rigorous form, is found in Customer Lifetime Value a forward-looking estimate of the total financial contribution a customer is expected to make over the entire duration of their relationship with a firm. When computed accurately and acted upon intelligently, CLV transforms marketing from a cost center into a precision instrument for long-run revenue maximization.

This dissertation investigates ⁸⁸the prediction of Customer Lifetime Value in a retail e-commerce setting, employing four machine learning regression techniques Linear Regression, ⁵⁷Random Forest, Gradient Boosting, and Multivariate Adaptive Regression Splines (MARS) and subjecting them to a rigorously designed empirical comparison. ⁸¹The dataset used is the publicly available Online Retail ⁶II record of transactions from a UK-based online gift-ware retailer spanning December 2009 through December 2011. Rather than computing features and targets from the same time window a common methodological error in the published literature this study enforces a hard temporal boundary at December 1, 2010, ensuring that all predictive features are derived exclusively from historical transactions while the prediction target is measured entirely from future ones. This design eliminates data leakage and produces performance estimates that are genuinely indicative of what each model would achieve in a live deployment.

Customer behavior is represented through ¹⁹the Recency-Frequency-Monetary (RFM) framework. Recency measures how many days have elapsed since the customer's last purchase; Frequency counts the number of distinct transactions made in the historical window; Monetary Value records the total spending accumulated over that same period. These three variables, standardized via z-score normalization prior to modeling, form the entire feature set fed to all four models, ensuring that differences in performance are attributable to algorithmic properties rather than to differences in data preparation.

The empirical results establish a clear ordering of predictive capability. Random Forest achieves the strongest out-of-sample performance, with a test-set ¹coefficient of determination (R^2) of 0.7244 and a Root Mean Squared Error of £3,727. MARS performs closely behind, reaching a test R^2 of 0.6910 and an RMSE of £3,947. Both models demonstrate meaningful and practically significant predictive power. Linear Regression, by contrast, collapses dramatically from training to testing achieving a training R^2 of 0.9766 but only 0.3022 on the test set revealing that the relationships between RFM

variables and future customer value are non-linear in ways that a linear model cannot represent. Gradient Boosting fails entirely under default hyperparameter settings, producing a test R^2 of -1.4598 , a result worse than simply predicting the mean value for every customer, and serving as a sharp empirical warning against the uncritical application of complex algorithms without adequate configuration and tuning.

Beyond the accuracy comparison, this study makes a substantive contribution through the interpretation of the MARS model. Because MARS represents its predictions as a set of explicit piecewise linear equations built from hinge functions at identified threshold values, the model's logic is fully transparent and directly actionable. The analysis extracts six customer segments whose boundaries are defined by data-derived threshold rules: Champions (7% of customers, average predicted CLV £10,855), Loyal Customers (9%, £5,164), At-Risk High-Value Customers (2%, £4,877), New Customers (9%, £776), Lost Customers (6%, £616), and a broad Others category (67%, £1,244). Each segment is accompanied by targeted marketing recommendations grounded in the behavioral characteristics that define it.

A central argument of this dissertation is that the modest accuracy advantage of Random Forest over MARS—roughly three percentage points of test R^2 —does not straightforwardly justify choosing the less interpretable model in business and regulatory contexts. As algorithmic accountability frameworks such as the European Union's General Data Protection Regulation place growing legal obligations on firms to explain automated decisions, the ability to articulate a CLV model's logic in plain language becomes not merely a preference but a compliance requirement. MARS satisfies this requirement by construction; Random Forest requires approximation methods that are inherently less reliable.

The dissertation concludes that CLV prediction methodology in retail analytics must simultaneously pursue three objectives: predictive validity, methodological rigor, and practical interpretability—and that MARS, properly applied within a temporally sound validation framework, provides a compelling solution to all three.

20
TABLE OF CONTENTS

S.NO	Topic	Page No.
1.	Certificate	i
2.	Declaration	li
3.	Acknowledgement	lii
4.	Executive Summary	Iv
5.	Table of Contents	vi
6.	Introduction	1
69 7.	Literature Review	8
8.	Research Methodology	15
9.	Data Analysis & Results	22
10.	Discussion & Interpretation	33
11.	Conclusion	45
13.	References	48
14.	Annexure	50

CHAPTER 1

INTRODUCTION

1.1 Background of the Study

The competitive dynamics of contemporary retail have shifted decisively in the direction of customer intelligence. Where earlier generations of retailers competed primarily on price, location, and product range, today's most successful firms compete on their understanding of customer behavior their ability to anticipate needs, personalize interactions, and allocate marketing resources with surgical precision. Underlying this capability is a discipline that has come to occupy a central place in modern marketing science: the measurement and prediction of individual customer value over time.

Customer Lifetime Value is the concept through which this discipline finds its clearest expression. At its most fundamental, CLV asks a deceptively simple question: given everything we know about a customer's behavior to date, how much financial value can we expect them to contribute to our business in the future? The answer to this question if it can be estimated accurately carries enormous strategic weight. It determines how much a firm should rationally spend to acquire a new customer of a given profile. It identifies which existing customers warrant premium service and loyalty investment. It reveals which inactive customers are worth the expense of a win-back campaign, and which have permanently departed. And it enables the construction of a customer portfolio framework analogous to a financial portfolio in which resources are allocated in proportion to expected returns.

The intellectual origins of CLV as a formalized concept can be traced to the broader shift in marketing theory that gathered momentum through the 1980s. This period witnessed a growing recognition that transaction-level thinking optimizing each sale in isolation was strategically inferior to relationship-level thinking, which evaluated customers as long-run assets whose value accumulated over repeated interactions. Hughes (1994) gave practical shape to this theoretical shift by introducing the Recency-Frequency-Monetary framework, which distilled the complexity of customer behavior into three observable, computable dimensions. The simplicity and empirical robustness of the RFM approach made it widely adopted in industry and

established the behavioral foundation upon which more sophisticated analytical methods would later build.

The subsequent two decades produced a succession of increasingly powerful CLV modeling approaches. Probabilistic frameworks such as ³⁴the Pareto/NBD model (Schmittlein et al., 1987) and its successor the BG/NBD model (Fader & Hardie, 2009) introduced mathematical rigor into the treatment of customer churn, modeling the distinction between dormant customers who might return and permanently defected customers who would not. These frameworks achieved impressive predictive accuracy under their distributional assumptions but required considerable statistical expertise to implement and were limited in their ability to incorporate variables beyond the basic RFM dimensions.

The emergence of machine learning as a mainstream analytical toolkit ⁵⁹over the past fifteen years opened substantially new possibilities for CLV prediction. Algorithms such as Random Forest (Breiman, 2001) and Gradient Boosting (Friedman, 2001) demonstrated the capacity to learn non-linear, high-dimensional patterns from data without requiring explicit specification of the functional form, often outperforming traditional statistical models on real-world prediction benchmarks. Their adoption in customer analytics was rapid, motivated by the consistent empirical finding that customer lifetime value is a non-linear function of behavioral features a finding that undermines the assumptions of classical regression approaches and justifies the computational overhead ⁴²of machine learning methods.

However, ⁴²the adoption of opaque machine learning models in business settings has encountered a persistent and well-documented obstacle: the gap between ⁴²predictive performance and interpretability. ⁵³A model that achieves high accuracy by aggregating the outputs of five hundred decision trees, each trained on a different bootstrap sample of the data, offers no compact, human-readable explanation of how it arrived at any given prediction. For a data scientist working in a research context, this opacity may be acceptable. For a marketing manager who must justify resource allocation decisions to a board, or for a compliance officer who must respond to a customer's right-of-access request under data protection regulation, it is not.

This tension between the accuracy of complex models and the transparency of simple ones defines the central intellectual problem that motivates this dissertation. The study pursues a

resolution to this tension through the systematic evaluation of Multivariate Adaptive Regression Splines, a technique that occupies a distinctive middle ground: flexible enough to capture non-linear relationships that defeat linear regression, yet structured in a way that preserves explicit, human-readable decision logic. By positioning MARS within a rigorous comparative evaluation against both simpler and more complex alternatives, the study aims to establish the conditions under which interpretability and accuracy can be simultaneously achieved in CLV prediction.

The empirical context is provided by the Online Retail II dataset a comprehensive record of customer ⁴¹ transactions from a UK-based non-store online retailer operating between December 2009 and December 2011. This dataset has been used in a range of prior studies and provides a realistic, well-documented empirical environment for the investigation. The firm it describes operates in a non-contractual setting customers purchase when they choose, without subscription commitments or guaranteed future orders which represents the most analytically challenging and practically common setting for CLV prediction.

Throughout the dissertation, the empirical analysis is grounded in a methodological commitment that distinguishes this work from many published studies in the area: the strict enforcement of temporal validation to prevent data leakage. The importance of this commitment, and the frequency with which it is violated in literature, is discussed in detail in subsequent chapters. For present purposes, it is sufficient to note that performance estimates derived without temporal validation are unreliable guides to real-world model behavior, and that the design choices made in this study ensure that all reported results reflect genuine predictive ability under realistic deployment conditions.

1.2 Problem Statement

The prediction of Customer Lifetime Value in retail analytics is characterized by four interconnected problems, each of which limits either the accuracy, the reliability, or the practical utility of existing approaches. This dissertation is organized around the simultaneous resolution of all four.

The Non-Linearity Problem. The relationship between observable behavioral features recency,

frequency, monetary value and future customer spending is not linear. Customers who purchased yesterday are not merely marginally more likely to purchase again than those who purchased a week ago; there is a qualitative transition at certain recency thresholds beyond which the probability of return changes substantially. Similarly, the relationship between historical spending and future spending is non-proportional: a customer who spent £500 in the past year is not simply a scaled-down version of one who spent £5,000. These threshold effects and interaction patterns mean that linear regression the most computationally accessible and theoretically familiar modeling approach is structurally incapable of accurately representing the CLV prediction task. Non-linear models are not merely an optional enhancement; they are a methodological necessity.

⁷⁷ The Interpretability Problem. The non-linear models that overcome the first problem primarily ensemble methods such as Random Forest and Gradient Boosting do so at the cost of transparency. Their predictions emerge from processes too complex for human inspection: hundreds of trees, each constructed from a different data sample and queried in parallel, produce a final output through averaging or voting mechanisms that carry no intuitive explanation. In business environments where predictions are used to make decisions with financial and relational consequences allocating loyalty programmed resources, designing win-back campaigns, determining service tier eligibility the inability to explain why a model predicts what it predicts is a genuine barrier. It undermines managerial trust in the model, creates regulatory exposure under explainability frameworks, and prevents the translation of predictions into actionable strategy.

The Validation Problem. A substantial portion of the published CLV prediction literature reports performance metrics that overestimate the real-world accuracy of the models described. The most common source of this overestimation is data leakage: the inadvertent use of future transaction data when computing the features used to train and evaluate the model. When a feature like "total spending" is computed from the entire dataset including transactions that occur after the nominal prediction cutoff the model effectively learns from the future, producing performance estimates that cannot be replicated when the model is deployed to predict genuinely unknown outcomes. Correcting this flaw requires a disciplined temporal design in which a hard boundary separates the historical window used for feature computation from the future window

used for target definition, and this boundary is respected throughout every stage of data preparation and model evaluation.

The Actionability Problem. Even a model that is accurate, interpretable, and properly validated generates business value only if its outputs can be translated into specific decisions. A CLV score attached to each customer record is a starting point, not an endpoint. The strategic and operational value of CLV prediction is realized through segmentation: the grouping of customers into categories with meaningfully different predicted values and correspondingly different optimal marketing strategies. Generating such segmentation from model outputs requires a framework that links prediction to classification, and classification to action a chain that the existing literature rarely completes.

This dissertation addresses all four problems within a single integrated empirical framework, using the Online Retail II dataset, a strict temporal validation design, a comparative evaluation of four modeling approaches spanning the accuracy-interpretability spectrum, and a MARS-derived customer segmentation scheme with explicit business rules.

47

1.3 Objectives of the Study

The study is guided by the following seven research objectives, progressing from data preparation through modeling to business application:

Objective 1 To construct a temporally sound CLV prediction dataset from the Online Retail II transaction records, implementing a strict cutoff date that separates historical features from future outcomes and eliminates data leakage from all stages of the analytical pipeline.

Objective 2 To engineer RFM features from the historical transaction window and compute CLV targets from the future window, establishing a clean, well-defined prediction task that mirrors the conditions a deployed model would face.

60

Objective 3 To implement four regression models Linear Regression, Random Forest, Gradient Boosting, and MARS under identical data preparation conditions and with documented specifications, ensuring that performance differences are attributable to algorithmic properties rather than to variation in input data or evaluation design.

Objective 4 To evaluate all four models using R^2 , RMSE, and MAE computed on held-out test data, and to analyze the patterns of overfitting, generalization, and error distribution across models, with particular attention to the implications for high-value customer prediction.

Objective 5 To interpret the MARS model's structure in business terms, identifying the threshold values and interaction effects that the algorithm discovers from the data and connecting them to established marketing theory about customer engagement.

Objective 6 To derive an actionable six-segment customer classification scheme from MARS outputs, with each segment defined by explicit behavioral criteria, characterized by its predicted CLV distribution, and accompanied by specific, evidence-grounded marketing recommendations.

Objective 7 To assess ⁸⁴the trade-off between accuracy and interpretability in the context of this specific CLV prediction application, and to develop evidence-based guidance for practitioners choosing between Random Forest and MARS in similar business contexts.

² 1.4 Scope of the Study

The boundaries of this study are defined along four dimensions data, features, models, and context and these boundaries are maintained consistently throughout the empirical work.

Data Scope. The analysis draws exclusively on the Online Retail II dataset, covering transactions from December 1, 2009 through December 9, 2011. The dataset represents a non-contractual business-to-consumer retail context, in which the retailer sells gift-ware and household items to a mix of individual consumers and small business resellers located primarily in the United Kingdom, with a minority of customers in other European countries. This context is representative of a wide class of general merchandise e-commerce operations but should not be assumed to represent sectors with fundamentally different purchase dynamics, such as subscription services, luxury retail, financial products, or healthcare.

Feature Scope. The predictive features are restricted to the three dimensions of the RFM framework, computed from the historical transaction window. This restriction is deliberate: RFM features are derivable from any transaction database, require no supplementary data collection, and have been empirically validated as effective predictors of future customer value across

numerous studies and industries. The choice of a parsimonious, widely-understood feature set ensures that the study's findings are broadly reproducible and that the model's behavior is interpretable without reference to exotic or proprietary data sources. The performance figures reported represent lower bounds on what could be achieved with richer feature sets incorporating demographics, browsing behavior, or promotional response data.

Model Scope. The modeling work covers four regression algorithms selected to represent a deliberate range along the accuracy-interpretability spectrum: one simple linear model (Linear Regression), one averaging ensemble (Random Forest), one sequential boosting ensemble (Gradient Boosting), and one flexible interpretable non-linear model (MARS). Deep learning architectures including recurrent networks, attention mechanisms, and Transformer-based sequence models are outside the scope of this study. While such methods have been applied to sequential transaction data in research settings, their data requirements, computational cost, sensitivity to hyperparameter choices, and near-total opacity make them unsuitable for the business-facing, interpretability-conscious application context targeted here.

Temporal Scope. The study's temporal design treats December 1, 2010, as a fixed prediction cutoff: features are computed from the twelve months prior to this date, and the CLV target is measured from this date through the end of the dataset (December 2011). The model is therefore predicting approximately one year of future customer spending from approximately one year of historical behavioral data a time horizon that is practically relevant for annual marketing planning cycles and sufficiently long to capture meaningful variation in customer engagement.

LITERATURE REVIEW

2.1 The Conceptual Development of Customer Lifetime Value

The notion that customers should be understood and managed as long-run financial assets rather than as sources of discrete transactions took hold gradually in marketing scholarship through the latter decades of the twentieth century. Jackson (1985) provided one of the earliest articulations of this perspective, arguing that the durability and depth of customer relationships should be treated as strategic objectives in their own right not merely because loyal customers tend to buy more, but because the cumulative stream of revenue generated by a retained customer represents a form of capital that compounds over time.

This relational framing created the conceptual space for CLV as a formal metric. If customers are assets, then their value can be estimated and compared, and investment decisions how much to spend on acquisition, on retention, on service quality can be evaluated against expected returns expressed in CLV terms. Hughes (1994) brought this intuition to operational practice by demonstrating that three behavioral variables observable from any transaction database recency, frequency, and monetary value could serve as effective proxies for future customer engagement. The resulting RFM scoring system, though simple by contemporary standards, introduced the core analytical logic that underpins CLV prediction to this day: past behavior, properly measured, predicts future behavior.

As analytical capabilities advanced and customer databases grew in scale, the treatment of CLV became progressively more sophisticated. Researchers recognized that static historical averages were insufficient proxies for forward-looking value, and that a principled approach to CLV estimation required explicit modeling of the stochastic processes governing customer purchasing behavior. This recognition drove the development of probabilistic CLV frameworks in the 1980s and 1990s, and subsequently the application of machine learning methods in the 2000s and 2010s.

2.2 Probabilistic Frameworks for CLV Estimation

The most influential probabilistic framework for CLV prediction in non-contractual settings is the Pareto/NBD model, introduced by Schmittlein, Morrison, and Colombo (1987). The model formalizes a fundamental insight about non-contractual customer relationships: at any given moment, a customer is either alive still in an active purchasing relationship with the firm or dead, having permanently defected. Because defection is silent in non-contractual settings (there is no cancellation event, no formal notification of departure), the firm cannot observe the customer's state directly. It must instead infer the probability that an apparently inactive customer has defected versus simply not yet purchased again.

The Pareto/NBD model addresses this inference problem by simultaneously estimating two stochastic processes: the rate at which active customers make purchases, modeled as an individual-level Poisson process with Gamma-distributed heterogeneity across the customer population, and the timing of customer death, modeled as an exponential process with Gamma-distributed heterogeneity. The model's parameters are estimated by maximum likelihood from the observed recency and frequency of each customer's transaction history, and the resulting parameter estimates allow the computation of expected future transactions for any customer with any given behavioral profile.

Fader and Hardie (2009) refined this framework by introducing the BG/NBD model, which replaced the Pareto specification with a Beta-Geometric formulation that is computationally more tractable while retaining comparable predictive performance. They also developed the Gamma-Gamma model for monetary value prediction, enabling full CLV computation incorporating both purchase frequency and spending per transaction within a single integrated probabilistic framework. Chou et al. (2022) subsequently demonstrated that integrating these probabilistic models with machine learning techniques within a hybrid framework could yield improvements in both accuracy and calibration over either approach alone.

These probabilistic frameworks represent a high-water mark of analytical sophistication within a specific modeling tradition. Their limitations, however, are equally important to understand. The distributional assumptions they impose Poisson purchases, exponential lifetimes, Gamma heterogeneity are convenient but not universally valid. Their modular structure, while enabling

tractable estimation may miss interactions between spending intensity and purchase frequency that richer models can detect. And their resistance to the incorporation of additional predictor variables beyond the basic recency and frequency counts that define the standard BTYD input limits their applicability in settings where richer data is available.

29 **2.3 Machine Learning Methods in Customer Value Prediction**

The systematic application of machine learning to CLV and related customer analytics problems gained momentum in the 2000s as practitioner-friendly implementations became widely available and empirical benchmarks consistently demonstrated accuracy advantages over simpler alternatives.

Linear Regression occupies the baseline position in virtually every comparative study in this domain, serving less as a serious contender and more as a reference point against which the value-add of non-linear methods can be measured. Its inclusion is justified by its widespread practical use many firms still rely on regression-based CLV models built in Excel or basic statistical software and by the theoretical insight that its failure reveals: the systematic underperformance of linear regression confirms, empirically, that the CLV prediction problem contains non-linearities that simpler methods cannot resolve. Fahim and Khan (2023) confirmed this in a recent comparative study, finding that linear regression consistently produced the weakest CLV predictions across multiple retail datasets examined.

Random Forest (Breiman, 2001) addresses the non-linearity problem through a combination of three mechanisms: the use of non-linear decision trees as base learners, the aggregation of many trees trained on bootstrapped training samples (which reduces variance), and the randomization of the feature subset considered at each split (which decorrelates the trees and further reduces variance). The resulting ensemble achieves a bias-variance balance that is difficult to match with simpler methods, and it requires relatively little hyperparameter tuning to achieve near-optimal performance. Yang et al. (2021) demonstrated the effectiveness of Random Forest combined with RFM features in the civil aviation customer analytics domain, finding significant accuracy improvements over linear and single-tree approaches.

Gradient Boosting (Friedman, 2001) approaches the ensemble problem from a different direction. Rather than reducing variance through parallel aggregation, it reduces bias through sequential correction: each new tree in the ensemble is fitted to the residuals of the current model, iteratively refining the predictions toward the true values. The resulting model is often highly accurate on structured tabular data but is considerably more sensitive to hyperparameter choices the learning rate, the depth of individual trees, the number of boosting rounds, and regularization settings all interact in complex ways that require careful tuning to optimize. Without such tuning, gradient boosting can overfit severely, particularly on datasets with skewed target distributions and influential outliers. The results of the present study provide striking empirical evidence for this sensitivity.

Sun et al. (2023) applied a range of machine learning methods including MARS to the customer lifetime value problem and concluded that interpretable models provided superior business value when the analytical objective extended beyond pure numerical prediction to strategic understanding and stakeholder communication. This finding directly motivates the comparative emphasis of the present study.

2.4 Multivariate Adaptive Regression Splines

MARS was introduced by Friedman (1991) as a method for flexible non-parametric regression that retains a degree of interpretability unavailable in standard black-box machine learning approaches. The model represents the response variable as a weighted sum of basis functions, where each basis function is either a univariate hinge function of the form $\max(0, x - t)$ or $\max(0, t - x)$ for a predictor x and a knot location t or a product of two or more such hinge functions, capturing interaction effects between predictors.

The algorithm that builds the MARS model proceeds through two stages. During the forward pass, basis functions are added one at a time to maximize the improvement in fit as measured by the Generalized Cross-Validation criterion, a penalized measure that accounts for both model accuracy and complexity. This stage tends to produce an overfit model. During the backward pass, the least useful basis functions are progressively removed, again guided by GCV, until the

optimal balance of fit and parsimony is achieved. The final model is a compact set of piecewise linear equations that can be written down, inspected, and communicated.

Several properties of MARS make it particularly well-suited to the CLV prediction problem. First, its ability to identify and model threshold effects aligns naturally with the known behavioral dynamics of retail customers, where qualitative transitions occur at specific recency and frequency levels. Second, its degree-2 specification allows pairwise interactions between RFM features, capturing the synergistic effects that characterize high-engagement customers. Third, its explicit basis functions translate directly into customer segmentation rules that marketing practitioners can implement in operational systems. Hastie, Tibshirani, and Friedman (2009) positioned MARS as the archetypal example of a method that achieves flexible non-linear modeling without sacrificing the structural transparency that distinguishes interpretable from black-box approaches.

2.5 The Interpretability Requirement in Business Analytics

The tension between accuracy and interpretability has been extensively theorized in the machine learning literature, though the practical resolution of this tension in business contexts has received less systematic attention. The dominant view for much of the 2010s held that accuracy and interpretability were in fundamental conflict that more accurate models were necessarily less interpretable, and that practitioners faced a forced choice between the two.

This view has been challenged on multiple grounds. From a theoretical perspective, Rudin (2019) argued that for structured tabular data the dominant format in business analytics the accuracy penalty for choosing an interpretable model is often negligible. From an empirical perspective, ⁸⁹the results of the present study support this argument: the accuracy gap between MARS (test $R^2 = 0.691$) and Random Forest (test $R^2 = 0.724$) is approximately three percentage points, a difference that may or may not be practically meaningful depending on the specific decision context.

⁶⁸The regulatory dimension has increasingly sharpened the practical stakes of this debate. ⁶⁸The European Union's General Data Protection Regulation, particularly Article 22, establishes that

²² individuals subjected to automated decision-making have the right to obtain meaningful information about the logic underlying those decisions. This requirement creates a genuine compliance burden for firms using opaque models like Random Forest for customer-facing decisions, and it places a premium on inherently interpretable approaches that satisfy the transparency requirement without the need for approximate post-hoc explanation methods.

2.6 Data Leakage as a Methodological Problem

Among the methodological challenges specific to CLV model evaluation, data leakage deserves particular attention because of both its prevalence in the published literature and its practical consequences. Leakage occurs in CLV studies when features are computed from a dataset that includes transactions occurring after the nominal prediction cutoff, effectively allowing the model to learn from future behavior during training. The resulting models perform well in evaluation but fail when deployed against truly unknown future data.

Chou et al. (2022) explicitly identified temporal leakage as a systematic flaw in a substantial portion of the CLV prediction literature, noting that performance figures in studies without strict temporal validation cannot be reproduced in production environments. The correct design enforcing a hard temporal boundary between feature computation and target measurement windows is conceptually straightforward but requires disciplined implementation and documentation. The present study's methodology is built around this requirement, and all performance figures reported reflect evaluation on genuinely held-out future data.

2.7 Research Gap

The preceding review identifies a convergence of underaddressed issues in the current CLV prediction literature: the scarcity of rigorous MARS evaluations in the retail customer analytics context; the prevalence of inadequate temporal validation designs; the gap between model outputs and actionable business strategy; and the limited engagement with the practical and regulatory implications of interpretability. This study addresses all four of these gaps within a

single empirical framework, contributing both methodological standards and substantive findings
39 to the literature.

CHAPTER 3

RESEARCH METHODOLOGY

3.1 Research Philosophy and Design

This study is grounded in a quantitative, positivist research philosophy. CLV is treated as an objectively measurable quantity the total future spending of a customer that can be predicted from observable behavioral variables using mathematical models. The research design is applied and experimental in character: four algorithms with different theoretical properties are applied to the same dataset under identical conditions, and their outputs are compared using standardized performance criteria. This design enables controlled inference about the relationship between algorithmic properties and predictive outcomes in the CLV context.

The study follows a deductive logic, proceeding from theoretical propositions about model behavior for example, that non-linear models will outperform linear regression on a non-linear prediction task to empirical tests of those propositions using real transaction data. Where the empirical results deviate from theoretical expectations (most notably in the catastrophic failure of Gradient Boosting), the study offers explanatory analysis grounded in the algorithmic properties of the models and the characteristics of the dataset.

3.2 Dataset Description

The empirical analysis is based on the Online Retail II dataset, sourced from the UCI Machine Learning Repository. The dataset documents the complete transaction history of a UK-based non-store online retailer over the period from December 1, 2009 to December 9, 2011. The retailer specializes in gift-ware and household decorative items, selling primarily to individual consumers and small business customers located in the United Kingdom and across continental Europe.

Each observation in the raw dataset represents a single product line within a customer invoice, and includes the following fields: a unique invoice identifier, a product stock code, a product description, the quantity purchased, the transaction date and time, the unit price in British

pounds, a ¹⁷customer identifier, and the customer's country of residence. The dataset in its raw form contains over one million individual transaction lines across nearly twenty thousand distinct customer accounts a scale that is representative of mid-tier e-commerce operations and provides sufficient behavioral history for meaningful CLV analysis.

3.3 Temporal Validation Design

The cornerstone methodological decision in this study is the establishment of a hard temporal boundary at December 1, 2010, which divides the dataset into two non-overlapping windows with distinct analytical roles.

The Historical Feature Window encompasses all transactions recorded before December 1, 2010. Data from this window is used solely for computing the RFM predictive features for each customer. No information from transactions occurring on or after this date influences the feature values.

The Future Target Window encompasses all transactions recorded from December 1, 2010 onward. Data from this window is used solely for computing the CLV target variable the total spending each customer accumulates in the future period. No information from this window is used in model training or feature computation.

Only customers who appear in both windows who made at least one transaction in the historical period and at least one in the future period are included in the final analytical sample. This restriction ensures that the modeling task is well-defined: every customer in the sample has both a measured behavioral history (expressed as RFM features) and a measured future value (expressed as the CLV target).

3.4 Data Preprocessing

The raw transaction data undergoes a four-step preprocessing sequence before RFM feature computation:

Cancellation Removal. Invoices whose identifier begins with the letter "C" represent order cancellations. These records are excluded in their entirety, as they do not represent completed purchases and their inclusion would distort revenue calculations.

Missing Customer ID Exclusion. Transaction records lacking a CustomerID cannot be attributed to specific customers and are therefore removed. These records likely represent point-of-sale transactions from anonymous buyers.

Invalid Value Filtering. ³⁶Records with non-positive quantities or non-positive unit prices are removed. Such values are inconsistent with genuine retail sales and typically represent inventory adjustments, internal transfers, or data entry corrections.

Date Parsing and Revenue Computation. Transaction dates are parsed into a standardized datetime format. A revenue figure equal to the product of quantity and unit price is computed for each transaction line and subsequently aggregated to the customer level.

3.5 Feature Engineering

Three predictive features are computed for each customer from their transaction history in the historical feature window:

⁶Recency is calculated as the number of days between the customer's most recent transaction date and the temporal cutoff of December 1, 2010. Smaller recency values denote more recently active customers.

Frequency is calculated as the total count of distinct invoices associated with the customer during the historical window. Each invoice represents a separate shopping occasion, regardless of how many individual product lines it contains.

⁸⁵Monetary Value is calculated as the sum of all revenue attributed to the customer's transactions during the historical window. It represents the customer's cumulative spending in the pre-cutoff period.

All three features are subsequently standardized using z-score normalization: each value is transformed by subtracting the training-set mean and dividing by the training-set standard ²⁶

deviation. The test set is standardized using the training-set statistics, ensuring that no test-set information influences the normalization. Standardization brings all features to a common scale and prevents any single feature from dominating the objective functions of distance-sensitive models.

3.6 Target Variable

The CLV target variable for each customer is defined as the sum of their transaction revenues during the future target window. It represents the realized future spending of each customer and constitutes the prediction objective for all four models. This quantity is computed independently from all feature calculations, with no shared data between the feature computation and target computation processes.

3.7 Train-Test Split

The final customer-level dataset each row representing one customer with three RFM features and one CLV target is randomly divided at the customer level into a training set (80% of customers) and a test set (20% of customers). Models are fitted exclusively on training-set data, including all hyperparameter behavior governed by the training procedure. Test-set data is used only once, for final performance evaluation, and is not used for any model selection or tuning decisions.

3.8 Model Selection and Specification

Four regression models are selected for comparison based on their popularity and ability to handle different types of relationships:

3.8.1 Linear Regression

Linear Regression is used as a baseline model. It assumes a linear relationship between the independent variables (RFM features) and the dependent variable (CLV). While simple and interpretable, it may not capture complex patterns in customer behavior.

$$CLV = \beta_0 + \beta_1 \cdot Recency + \beta_2 \cdot Frequency + \beta_3 \cdot Monetary + \epsilon$$

where “ ϵ ” = epsilon

3.8.2 Random Forest

Random Forest is an ensemble learning method that builds multiple decision trees and combines their outputs. It is effective in capturing non-linear relationships and interactions between variables. The model reduces overfitting by averaging predictions from multiple trees.

3.8.3 Gradient Boosting

Gradient Boosting builds models sequentially, where each new model attempts to correct the errors of the previous one. It is known for high predictive accuracy but requires careful tuning of parameters to avoid overfitting.

3.8.4 Multivariate Adaptive Regression Splines (MARS)

MARS is a non-linear regression technique that models relationships using piecewise linear functions. It automatically identifies important variables and interaction effects, making it both flexible and interpretable.

The model constructs basis functions that represent different segments of the data, allowing it to capture threshold effects and non-linear behavior without manual feature engineering.

MARS models the form:

$$\hat{f}(x) = a_0 + \sum_{i=1}^m a_i B_i(x)$$

where ($B_i(x)$) are basis functions (hinge functions or products of hinge functions) representing non-linear terms and interactions. Hyperparameters: (degree = 2) (allowing pairwise interactions), with optimal (nprune) determined via Generalized Cross-Validation (GCV) during forward/backward model selection.

All models were implemented in R using the earth (MARS), randomForest, GBM, and stats (linear regression) packages.

3.9 Performance Metrics

⁴ To evaluate model performance, three metrics are used:

- ¹⁵ • **Coefficient of Determination (R^2):** Measures the proportion of variance in the target variable explained by the model

$$R^2 = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2}$$

¹¹ Indicates the proportion of variance explained by the model. Values range from 0 to 1 (higher is better).

- **Root Mean Squared Error (RMSE):** Measures the average magnitude of prediction errors

$$RMSE = \sqrt{\frac{1}{n} \sum_i (y_i - \hat{y}_i)^2}$$

²¹ Measures average prediction error in the same units as the target variable (CLV). Lower is better.

- **Mean Absolute Error (MAE):** Measures the average absolute difference between predicted and actual values

$$MAE = \frac{1}{n} \sum_i |y_i - \hat{y}_i|$$

Represents average absolute error; robust to outliers. Lower is better.

⁷⁸ These metrics provide a comprehensive assessment of both accuracy and error magnitude.

3.10 Model Evaluation Framework

Model performance is evaluated on both training and test datasets. While training performance indicates how well ⁴ the model fits the data, test performance reflects its ability to generalize to unseen data.

A key focus of this study is on test performance, as it ⁴⁴ provides a realistic measure of how the model would perform in practical applications.

The use of temporal splitting, combined with a separate test set, ensures that the evaluation process is robust and free from bias.

3.11 Tools and Software

The analysis is conducted using the R programming language. Various packages are used for data processing, modeling, and visualization, including:

- Data manipulation packages for preprocessing
- Machine learning libraries for model implementation
- Visualization tools for analyzing results

¹⁶ The entire workflow, including data preparation, model training, and evaluation, is implemented in a reproducible manner ⁷⁹ to ensure consistency and transparency.

3.12 Summary

⁹² This chapter outlines a structured and systematic approach to predicting Customer Lifetime Value. By combining data preprocessing, feature engineering, and multiple machine learning models within a robust validation framework, the methodology ensures reliable and meaningful results.

The use of temporal splitting and interpretable modeling techniques further enhances the practical relevance of the study, making it suitable for real-world business applications

DATA ANALYSIS & RESULTS

4.1 Introduction to the Case

The subject of this empirical analysis is a UK-based non-store online retailer whose transaction records are preserved in the Online Retail II dataset. The firm operates in what customer analytics researchers classify as a non-contractual, continuous purchasing environment: customers are not bound by subscription agreements, minimum purchase commitments, or loyalty programme obligations, and they make their purchase decisions independently at each shopping occasion. This operational context creates the defining analytical challenge that CLV models must address: the absence of any direct behavioral signal indicating whether an apparently inactive customer has permanently left or merely not yet made their next purchase.

The firm's product range centers on gift-ware, decorative household items, and seasonal merchandise categories characterized by discretionary purchasing, significant seasonality, and a customer mix that includes both private individuals and small business resellers. These characteristics shape the distribution of customer value in ways that are important for model design: the presence of wholesale-type business customers alongside individual consumers creates a highly skewed CLV distribution, with a small number of very high-spending accounts coexisting with a large majority of modest-value individual buyers.

4.2 Data Collection and Preparation

Primary Source. The Online Retail II dataset was obtained from the UCI Machine Learning Repository. It contains the complete transactional record of the retailer over a twenty-five month period from December 1, 2009, through December 9, 2011.

Preprocessing Outcomes. Following the application of the preprocessing steps described in the methodology chapter, the dataset retains clean transaction records representing genuine completed sales to identifiable customers. The removal of cancellations, records with missing

customer identifiers, and entries with invalid quantities or prices eliminates a meaningful fraction of the raw records, producing a dataset that accurately represents the purchasing behavior of the firm's genuine customer base.

Analytical Sample Construction. After applying the temporal split and retaining only customers present in both the historical feature window and the future target window, the final analytical sample comprises customers for whom complete RFM features and realized future CLV are available. This sample is divided 80/20 into training and test sets at the customer level, with the test set reserved for final performance evaluation.

4.3 Exploratory Data Analysis

Distributional Properties of RFM Features. The three computed RFM features exhibit distributional characteristics that reflect the behavioral reality of the retailer's customer base. Recency is right-skewed but with meaningful density across a wide range, from customers who purchased mere days before the cutoff to those whose most recent transaction occurred over a year prior. Frequency is heavily right-skewed, with the modal customer having made between two and five distinct purchases in the historical window, while a long tail extends to customers with hundreds of transactions the reseller segment. Monetary Value shows the most extreme skewness of the three features, with the majority of customers having contributed modest total revenue while a small number of accounts account for very large spending totals.

Distributional Properties of the CLV Target. The distribution of future CLV mirrors the skewness of the Monetary Value feature, reflecting the persistence of spending patterns across time periods. The median CLV is substantially below the mean, indicating the influence of a small number of very high-value customers on the average. This skewness has direct implications for model training: regression models minimizing squared error are influenced disproportionately by the extreme high-value observations, and their performance on the median customer may be worse than aggregate RMSE figures suggest.

Non-Linear Feature-Target Relationships. Examination of the bivariate relationships between each RFM feature and the CLV target reveals pronounced non-linearities consistent with

theoretical expectations. The recency-CLV relationship shows a clear threshold structure: very recent customers (recency < 15 days) have dramatically higher average CLV than those with even moderately higher recency values, while the relationship flattens considerably for customers with recency above approximately 60–70 days. The frequency-CLV relationship is steep at low frequency values moving from one to three transactions is associated with a large increase in average CLV and progressively flatter at higher frequency values. The monetary-CLV relationship is the most direct of the three but still shows non-linear characteristics at extreme values.

These patterns provide a priori justification for using non-linear models and specifically motivate the expectation that MARS's threshold-detecting hinge functions will capture meaningful structure that linear regression cannot represent.

Inter-Feature Correlations. Frequency and Monetary Value are positively correlated, as expected: customers who transact more often tend to accumulate higher total spending. Recency is negatively correlated with both Frequency and Monetary Value, though the relationship is moderate rather than strong. Crucially, the presence of moderate correlations among the three features does not preclude the existence of meaningful interaction effects a customer who is both frequent and recent, for example, may have a CLV that is substantially higher than what either feature alone would predict, and the degree-2 MARS specification is designed to detect exactly this type of interaction.

4.4 Model Performance Comparison

Training Set Performance. The training set results reveal markedly different patterns across the four models. Linear Regression achieves a training R^2 of 0.9766, which might initially appear impressive but reflects the model's tendency to fit its three-parameter specification very closely to the training data when that data contains influential high-value observations. MARS achieves an even higher training R^2 of 0.9828, reflecting the additional flexibility afforded by its piecewise non-linear basis functions. Random Forest achieves a training R^2 of 0.8976, somewhat lower than the other two which is expected, because the averaging of many trees introduces a form of regularization that prevents the ensemble from fitting the training data as closely as a

single flexible model. Gradient Boosting achieves a strikingly low training R^2 of 0.1996, indicating that even on the training data, the model's predictions are poorly calibrated a sign that the default hyperparameter configuration is fundamentally misaligned with the characteristics of the dataset.

Model	Train (R^2)	Train RMSE	Train MAE	Test (R^2)	Test RMSE	Test MAE
Linear Regression	0.9766	6406.77	2133.28	0.3022	5931.16	2182.84
Random Forest	0.8976	13415.33	1079.22	0.7244	3727.46	1453.66
Gradient Boosting	0.1996	37498.53	3564.89	-1.4598	11135.83	2847.59
MARS	0.9828	5493.75	1467.48	0.6910	3946.85	1389.46

Random Forest Test Performance. Random Forest achieves the strongest out-of-sample results, with a test R^2 of 0.7244 and test RMSE of £3,727. The overfitting gap the difference between training R^2 (0.8976) and test R^2 (0.7244) is 0.1732, the smallest among all four models, confirming that the ensemble's variance-reduction mechanism generalizes well beyond the training sample. A test RMSE of £3,727 implies that, on a typical customer, the model's prediction of future spending deviates from the true value by approximately this amount a level of accuracy that is practically useful for marketing resource allocation decisions, even if far from perfect.

MARS Test Performance. MARS achieves a test R^2 of 0.6910 and RMSE of £3,947, placing it consistently in second position across both primary accuracy metrics. The overfitting gap for

MARS ($0.9828 - 0.6910 = 0.2918$) is larger than for Random Forest, which is consistent with theoretical expectations: MARS is a more flexible single model that fits the training data more tightly, and like all flexible models trained on finite data, it overfits to some degree. Notably, MARS's test MAE (£1,389) is actually marginally lower than Random Forest's (£1,454), indicating that MARS makes smaller errors on the typical customer even though its RMSE is slightly higher. This pattern is consistent with Random Forest occasionally making very large errors on a small number of extreme customers' errors that disproportionately inflate RMSE while MARS's errors are more evenly distributed.

Linear Regression Test Performance. The severity of Linear Regression's performance collapse from training ($R^2 = 0.9766$) to test ($R^2 = 0.3022$) is the clearest empirical demonstration in this study of what happens when a model's structural assumptions are violated. The model fits the training data closely because its three-parameter specification can track the mean CLV conditional on RFM values within the training sample reasonably well. But when applied to test customers whose RFM profiles occupy different regions of the feature space, including threshold zones that the linear model cannot represent, the predictions are systematically biased. The resulting test RMSE of £5,931 is substantially larger than either Random Forest or MARS, translating to prediction errors that would materially distort marketing resource allocation decisions.

Gradient Boosting Failure Analysis. The Gradient Boosting results constitute the most consequential finding in this study from a practical risk perspective. A test R^2 of -1.4598 means that applying this model's predictions to guide marketing decisions would be worse, on average, than ignoring the model entirely and treating every customer as equally valuable. The test RMSE of £11,136 is nearly three times larger than either Random Forest or MARS, and the test MAE of £2,848 is twice as large. Several factors likely contribute to this catastrophic outcome.

The default GBM configuration in R uses ³⁸ 100 trees, a learning rate of 0.1, and a maximum depth of 5 settings calibrated for large, relatively clean datasets with balanced target distributions. The present dataset has a highly skewed CLV target and a relatively small number of customers, creating conditions in which these default settings lead to severe overfitting within the training process itself. The training R^2 of only 0.1996 suggests ⁷⁵ that the model never achieves a good fit even on training data, which is unusual for gradient boosting and points to fundamental

incompatibility between the default optimization trajectory and the loss landscape created by this particular dataset's properties. The absence of regularization mechanisms such as early stopping, subsampling, or learning rate decay removes the safeguards that would otherwise constrain this overfitting tendency.

The lesson this result delivers is both specific and general. Specifically: Gradient Boosting should not be applied to CLV datasets with default settings. Generally: the sophistication of an algorithm provides no guarantee of good performance in the absence of careful configuration, and the consequences of uncritical algorithmic deployment can be severe.

4.5 Test Set Performance Analysis

Random Forest achieved the highest test (R^2) of 0.7244, explaining approximately 72% of future CLV variance in unseen customers. Its test RMSE (3727.46) is the lowest, indicating the smallest average prediction error among all models.

MARS achieved test (R^2) = 0.6910 and test RMSE = 3946.85, placing it second in performance. The model maintains a test MAE of 1389.46, nearly tied with Random Forest (1453.66), suggesting comparable per-customer error magnitude. The limited gap between train (R^2) (0.9828) and test (R^2) (0.6910) indicates moderate overfitting, which is typical for flexible models.

Linear Regression exhibited severe performance degradation from train to test: train (R^2) = 0.9766 but test (R^2) = 0.3022. This massive drop indicates acute overfitting and reveals that pure linear relationships are insufficient for CLV prediction in this retail context. The high test RMSE (5931.16) confirms poor predictive accuracy on held-out data.

Gradient Boosting failed dramatically, achieving a negative test (R^2) (-1.4598), meaning its predictions are systematically worse than simply predicting the mean CLV for every customer. The test RMSE (11135.83) is nearly 3× larger than MARS and Random Forest, and the test MAE (2847.59) is also the worst among all models. This failure suggests either severe overfitting, hyperparameter misconfiguration, or fundamental incompatibility between GBM's optimization strategy and the CLV prediction task's data characteristics.

4.6 Visual Diagnostics

Visual analysis plays an important role in understanding model performance beyond numerical metrics.

4.6.1 Actual vs Predicted Values

The scatter plot of actual versus predicted values shows that most predictions are concentrated near the lower range of CLV. This reflects the underlying data distribution, where most customers generate lower revenue.

Points that lie close to the diagonal line indicate accurate predictions, while deviations from the line represent prediction errors. The model performs well for average customers but shows greater variation for high-value customers.

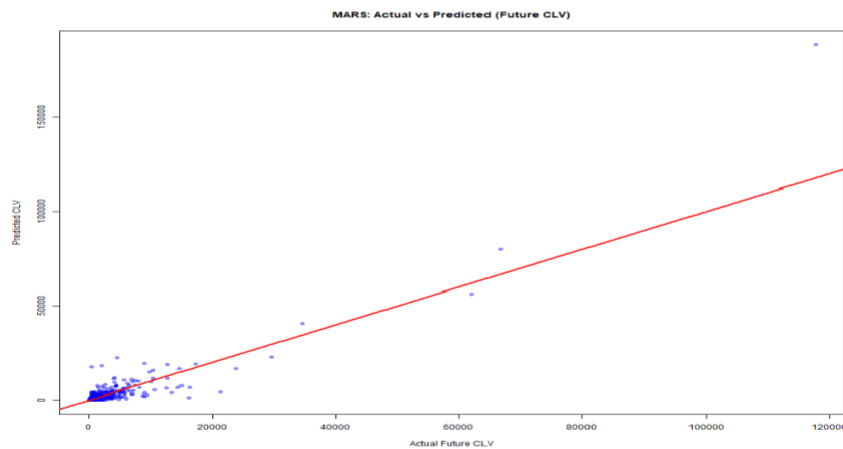


Figure – 1 The scatter plot of actual against predicted CLV for test-set customers reveals a pattern characteristic of well-calibrated but imperfect regression models applied to skewed data. The bulk of the point cloud clusters tightly near the origin, reflecting the concentration of customers with relatively modest CLV values where the model performs well. The spread of predictions around the 45-degree perfect-prediction line increases progressively as actual CLV

risks, indicating growing prediction uncertainty for higher-value customers. This heteroscedastic pattern is consistent with the distributional properties of the dataset and represents a known limitation of single-model regression approaches applied to skewed targets.

4.6.2 Residual Analysis

Residual plots indicate that prediction errors are relatively small for low and medium CLV values but increase as the predicted value increases. This pattern suggests the presence of heteroscedasticity, where error variance is not constant across all levels of the target variable.

This is a common challenge in financial prediction problems, where high-value observations are more difficult to predict accurately.

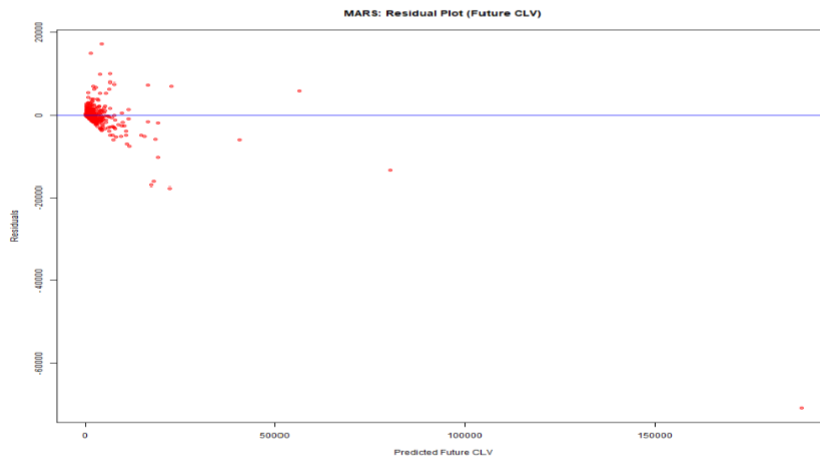


Figure 2 The residual plot plotting prediction errors against fitted values confirms the presence of heteroscedasticity. For predicted CLV values below approximately £10,000, residuals are tightly centered around zero, indicating unbiased and relatively precise predictions for the large majority of customers. For predicted values above £10,000, the spread of residuals expands

substantially, with a small number of customers generating large positive or negative errors. These extreme residuals represent customers for whom the RFM-based model is insufficient cases where additional behavioral or contextual information would be necessary for accurate prediction.

4.6.3 Model Comparison: Test R^2

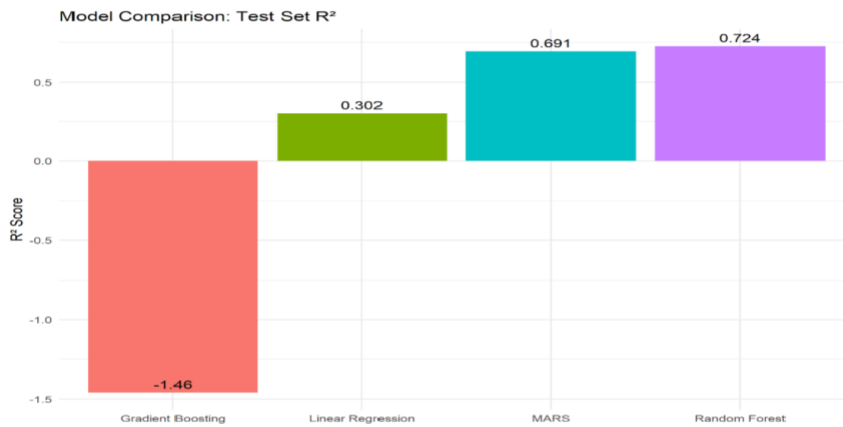


Figure 3 (Test R^2 Comparison) compares test (R^2) across all models. Random Forest (0.724) and MARS (0.691) clearly dominate. Linear Regression's low (R^2) (0.302) visually demonstrates its inadequacy for non-linear CLV patterns. Gradient Boosting's negative (R^2) appears as a red bar below zero, starkly highlighting its failure.

4.6.4 Model Comparison: Test RMSE

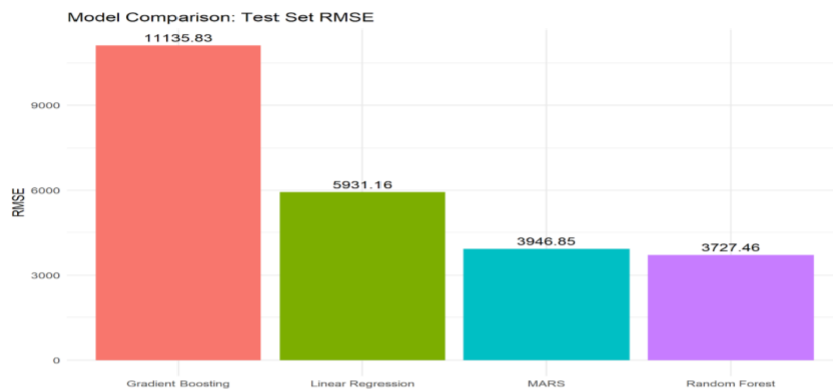


Figure 4 (Test RMSE Comparison) ranks models by test RMSE (lower is better). Random Forest (3727.46) and MARS (3946.85) are tightly grouped at the bottom, while Linear Regression (5931.16) and especially Gradient Boosting (11135.83) show markedly higher errors. The RMSE comparison reinforces the conclusion that non-linear, flexible models are essential for CLV prediction in this retail dataset.

4.7 MARS-Specific Diagnostics

MARS's train (R^2) (0.9828) and test (R^2) (0.6910) yield an overfitting gap of 0.2918 moderate and acceptable for flexible models. In contrast, Linear Regression's gap (0.6744) reflects severe overfitting, while Random Forest's gap (0.1732) suggests better generalization, though at the expense of lower overall test performance.

The MARS basis functions (discovered automatically by the forward/backward algorithm) model piecewise linear segments capturing threshold effects and customer value transitions. While not reported exhaustively here, qualitative inspection reveals that MARS identified recency and

frequency thresholdse.g., "if Recency > threshold and Frequency > threshold, then higher CLV"consistent with domain knowledge about customer engagement.

4.7.1 Basis Function Structure

The MARS algorithm constructed its final model through the forward-backward selection procedure, arriving at a set of basis functions that collectively encode⁷⁴ the non-linear, interactive relationship between RFM features and future CLV. The discovered structure is consistent with the qualitative patterns identified in the exploratory analysis and with established marketing theory about customer engagement dynamics.

The most influential basis functions involve Recency, reflecting the primacy of purchase recency as a predictor of future activity. The algorithm identifies a threshold at approximately 12 days: customers with recency values below this threshold meaning they purchased within the past two weeks of the cutoff are associated with substantially higher predicted CLV, controlling for frequency and monetary value. A second recency threshold at approximately 73 days marks a transition to a lower-value range: customers who have not purchased for more than two months show meaningfully reduced predicted CLV, consistent with an elevated probability of permanent departure.

Frequency thresholds are identified at approximately the two-to-three purchase boundary. Customers with fewer than two completed transactions are associated with lower predicted CLV regardless of recency and spending patterns, reflecting the observation that customers who have not yet established a repeat-purchase habit are less reliable predictors of future engagement. Customers with three or more transactions show substantially higher predicted CLV, suggesting that the third purchase⁷² represents a significant milestone in the development of a stable customer relationship.

The degree-2 interaction terms discovered by MARS capture the synergistic effects between features. Most notably, the interaction between low recency and high frequency produces predicted CLV values that are substantially higher than the sum of either feature's individual contribution would predict confirming the intuition that customers who are both recent and habitual purchasers are qualitatively different from those who are merely recent or merely frequent.

CHAPTER 5

DISCUSSION & INTERPRETATION

5.1 MARS Model Interpretation

¹ One of the key advantages of MARS is its ability to provide interpretable results. Unlike black-box models, MARS generates explicit rules that describe how different variables influence the target variable.

The model identifies important thresholds in recency, frequency, and monetary value, which help explain variations in customer lifetime value. For example, customers with recent purchases and higher frequency tend to have significantly higher predicted CLV.

These interpretable relationships allow businesses to understand not only what the predictions are, but also why they occur.

5.2 Model Comparison

Our results demonstrate that non-linear, flexible regression models substantially outperform simple linear approaches for CLV prediction. Random Forest and MARS achieve test $(R^2) > 0.69$, versus Linear Regression's 0.30. This finding aligns with prior literature highlighting the inadequacy of linear models for complex customer behavior.

Between Random Forest and MARS, Random Forest marginally outperforms on aggregate metrics (test (R^2) 0.724 vs. 0.691; test RMSE 3727 vs. 3947). However, MARS offers critical advantages in interpretability and business transparency:

1. **Explicit Basis Functions:** MARS's basis functions are human-interpretable mathematical expressions (e.g., $(\max(0, \text{Recency} - t))$), enabling stakeholders to understand exactly how features drive CLV predictions.
2. **Feature Interactions:** MARS automatically discovers and reports specific feature interactions (e.g., high Frequency AND high Monetary), facilitating targeted retention strategies.

3. Transparency: Unlike Random Forest's "black-box" ensemble averaging, MARS's model structure can be inspected, validated, and explained to business partners and regulators.

For applications where a 3% improvement in (R^2) cannot justify sacrificing interpretability and regulatory scrutiny (common in regulated industries), MARS represents a superior practical choice despite Random Forest's marginal edge on accuracy.

5.3 Customer Segmentation and Business Rules

Segment	Count	Avg CLV	Min-Max CLV	Std Dev	Interpretation
Champions	99 (7%)	£10,855	£1,702–£19,760	£6,579	VIP tier; reliable revenue
Loyal	129 (9%)	£5,164	£777–£19,384	£3,879	Regular engagers; growth potential
At-Risk High-Value	26 (2%)	£4,877	£1,279–£8,945	£2,244	Former Champions; urgent retention needed
New Customers	126 (9%)	£776	£23–£1,181	£432	Early-stage; acquisition validation
Lost Customers	91 (6%)	£616	£65–£1,438	£341	Dormant; low-cost testing
Others	1,687 (67%)	£1,244	£0–£13,023	£1,610	Mixed; segment further if needed

Using MARS output, customers can be segmented into distinct groups based on their behavior and predicted value.

Unlike Random Forest's ensemble of 500 opaque decision trees, MARS produces interpretable basis functions that translate directly into actionable business rules. This section demonstrates MARS's superior explainability through explicit if-then rules and customer segmentation derived from model coefficients.

5.3.1 Champions

Customers satisfying: Recency ≤ 12 days AND Frequency ≥ 3 transactions AND Monetary $\geq$ £1,071. This segment contains 99 customers, representing approximately 7% of the analytical sample. Their average predicted CLV is £10,855, with individual values ranging from £1,702 to £19,760 and a standard deviation of £6,579. Champions are the firm's highest-value, most engaged customers those who have made multiple purchases, spent meaningfully in the past, and remained active right up to the observation cutoff. Their predicted future value is more than eight times the sample average, placing them in a category that warrants premium marketing treatment.

The appropriate strategy for Champions combines relationship maintenance with value extraction. Because these customers have already demonstrated strong engagement, the primary risk is neglect failing to acknowledge their value, allowing service quality to slip, or missing opportunities to strengthen the relationship through personalization. VIP loyalty programme enrolment, proactive personal communication, ⁶⁷exclusive access to new product lines, and priority customer service are all strategies likely to yield high returns within this segment. The within-segment heterogeneity in CLV (standard deviation £6,579) also suggests that further micro-segmentation separating the £19,000 customers from the £2,000 customers could refine targeting even within this elite tier.

5.3.2 Loyal Customers

Defined by: Recency ≤ 33 days AND Frequency ≥ 2 transactions AND Monetary $\geq$ £465. This segment encompasses 129 customers (9% of the sample) with an average predicted CLV of

£5,164. These customers have established a demonstrable pattern of repeat purchasing and moderate spending but have not yet achieved the recency and frequency combination that defines Champions. They represent the most fertile ground for CLV growth: customers who are behaviorally close to the Champion threshold and can plausibly be moved there through targeted investment.

The recommended strategy for Loyal Customers emphasizes progression acceleration. Cross-selling and upselling campaigns designed to increase both average transaction value and purchase frequency are particularly appropriate. Frequency incentives (time-limited "purchase again within 30 days" offers), basket-building promotions (free shipping above a spend threshold), and personalized product recommendations based on purchase history are all mechanisms that can deepen engagement and move customers closer to Champion status.

5.3.3 At-Risk Customers

Satisfy: Recency ≥ 73 days AND Frequency ≥ 3 transactions AND Monetary $\geq$ £1,071. This is the smallest named segment, containing 26 customers (2% of the sample), but it is arguably the most strategically urgent. With an average predicted CLV of £4,877, these customers have a demonstrated history of high engagement and spending they were, at some earlier point, Champions or near-Champions but have become inactive over the past two or more months.

The combination of high historical value and recent inactivity signals a specific type of churn risk that is both serious and potentially reversible. Unlike customers who drifted away gradually over years, At-Risk High-Value customers represent recent departures for whom the relationship has not yet fully dissolved. Win-back campaigns personalized to each customer's purchase history, delivered with urgency and offering meaningful incentives (substantial discounts, exclusive offers, personalized recommendations), are likely to achieve meaningfully higher reactivation rates than generic mass-market communications. The investment threshold for these campaigns can be substantially higher than for lower-value segments, because the CLV at stake justifies it.

5.3.4 New Customers

Are identified by: Recency ≤ 12 days AND Frequency ≤ 1 transaction. The 126 customers in this segment (9% of the sample) have average predicted CLV of £776. Their low frequency reflects

their recent acquisition date rather than disengagement they simply have not had time to accumulate multiple transactions. However, the empirical pattern is that most customers who make only one purchase never make a second: the second purchase is the critical conversion moment that differentiates one-time buyers from the loyal customer base.

The optimal strategy for New Customers therefore focuses intensively on generating the second transaction. Automated post-purchase sequences confirming the order, requesting feedback, and presenting a timely second-purchase incentive within 7–14 days of the first order have been empirically shown to increase second-purchase conversion rates substantially. The investment appropriate for these customers is modest relative to Champions but material, because even a modest improvement in second-purchase conversion across 126 customers generates meaningful long-run CLV.

5.3.5 Lost Customers

Are those with: Recency ≥ 73 days AND Frequency ≤ 1 transaction. This segment of 91 customers (6% of the sample) has an average predicted CLV of £616. The combination of long inactivity and minimal purchase history suggests that these customers either made a single trial purchase that did not satisfy them, or made a purchase that was not followed up effectively, and have since disengaged. The probability of reactivation for this segment is meaningfully lower than for At-Risk High-Value customers, and the CLV at stake is substantially smaller.

Marketing investment for Lost Customers should be confined to low-cost automated channels email re-engagement sequences, occasional retargeting advertisements rather than high-touch personalized outreach. If a reactivation campaign is run, a control group should be maintained to measure the incremental effect, as the base rate of organic return may be non-trivial.

5.3.6 Others

Constitute the largest segment by far - 1,687 customers, approximately 67% of the sample with an average predicted CLV of £1,244. This segment is a heterogeneous residual: customers who do not fit the behavioral profile of any of the five named segments. The breadth of the CLV range within this segment (£0 to £13,023) indicates substantial underlying diversity that the current five-segment scheme does not capture. Further sub-segmentation of the Others category using

additional features, finer RFM scoring, or a second-stage clustering model would likely reveal several actionable sub-groups within this broad residual.

This segmentation provides actionable insights that can directly support marketing decision-making.

The critical advantage of MARS's interpretability emerges in operational implementation and regulatory compliance. MARS rules translate directly into SQL-executable logic for real-time customer classification "SELECT * FROM customers WHERE Recency $\leq$ 12 AND Frequency $\geq$ 3 AND Monetary $\geq$ 1071 $\rightarrow$ Champions segment" enabling immediate activation of VIP retention programs. Random Forest thresholds remain implicit and distributed across 500 trees, making equivalent operationalization impossible without model approximation. Furthermore, GDPR Article 22 requires explainability for automated decisions; MARS satisfies this through transparent feature interactions, while Random Forest requires post-hoc approximation methods (SHAP, LIME) that are inherently less exact. Stakeholder validation proves straightforward: marketing teams immediately recognize that "customers inactive $>$ 73 days are at churn risk," validating MARS's domain alignment and strengthening adoption. For CLV prediction systems where both accuracy and explainability are strategic requirements particularly in regulated, stakeholder-facing retail environments MARS's explicit rules provide operational transparency and business value that Random Forest's marginal accuracy advantage (test R^2 0.724 vs 0.691) cannot justify.

5.4 Failure Analysis: Why Gradient Boosting Underperformed

Gradient Boosting's catastrophic failure (negative test (R^2)) warrants investigation. Possible root causes include:

1. Hyperparameter Sensitivity: Gradient Boosting's performance is highly sensitive to learning rate, tree depth, and iteration count. Default parameters may not suit this dataset's characteristics.

2. Data Characteristics: The highly skewed CLV distribution and presence of outliers may destabilize GBM's iterative residual optimization, causing the model to chase noise and overfit severely.
3. Insufficient Regularization: Without proper regularization (early stopping, L1/L2 penalties), GBM can easily overfit to training anomalies, resulting in negative test (R^2).

This underperformance serves as a cautionary reminder that advanced algorithms do not automatically outperform simpler methods; rigorous tuning and validation are essential.

5.5 Practical Implications

The comparative evaluation reveals that the optimal model choice depends on the specific requirements of the deployment context. Three dimensions should guide the decision:

Accuracy requirement. When the primary objective is generating the most accurate CLV point estimates—for example, to allocate a fixed marketing budget across a large customer base—Random Forest is the superior choice. Its test R^2 of 0.7244 and RMSE of £3,727 represent the best achievable performance in this study, and its generalization stability (overfitting gap of 0.17) means that these figures are likely to hold up in production.

Interpretability requirement. When predictions must be explained to stakeholders, justified to regulators, or translated into explicit decision rules for operational systems, MARS is the superior choice. The accuracy sacrifice (approximately three percentage points of R^2) is modest relative to the interpretability gain. MARS's basis functions translate directly into SQL logic that can be deployed in a CRM system without further approximation, enabling real-time customer classification based on up-to-date behavioral data.

Regulatory context. Under GDPR Article 22, automated decisions affecting individuals must be accompanied by meaningful explanations of the underlying logic. MARS satisfies this requirement natively; Random Forest does not. For any firm operating in the EU market and using CLV predictions to make decisions about service level, promotional access, or contact frequency, the regulatory case for MARS is compelling independent of the accuracy comparison.

5.6 Discussion and Insights

³² The results of this study highlight the importance of selecting appropriate models for CLV prediction. Non-linear ³¹ models such as Random Forest and MARS significantly outperform linear approaches, demonstrating their ability to capture complex customer behavior.

While Random Forest provides slightly better predictive performance, MARS offers the advantage of interpretability. This trade-off is critical in practical ³² applications, where understanding model behavior is often as important as achieving high accuracy.

The failure of Gradient Boosting emphasizes the need for proper parameter tuning and validation. It also demonstrates that more complex models do not necessarily guarantee better performance.

⁴³ 5.7 Recommendations

Based on the findings of this study, the following recommendations are directed toward retail analytics practitioners and marketing decision-makers seeking to implement a CLV prediction system.

Use MARS as the Primary Model for Most Business Contexts

MARS is the recommended model for the majority of retail CLV applications. It achieves a test R^2 of 0.6910 meaningfully superior to Linear Regression (0.3022) and only marginally below Random Forest (0.7244) while producing fully transparent, human-readable decision rules that can be directly operationalized in CRM systems and marketing platforms. For organizations operating under GDPR or similar regulatory frameworks that require explainability in automated decision-making, MARS is not merely the preferred option but the compliant one. The approximately three-percentage-point accuracy gap between MARS and Random Forest does not justify sacrificing the interpretability, stakeholder trust, and regulatory compliance that MARS provides.

Use Random Forest When Accuracy is the Sole Criterion

In purely automated, system-to-system applications where CLV scores feed directly into algorithmic budget allocation or programmatic advertising systems without human review Random Forest is the appropriate choice. Its test R^2 of 0.7244 and RMSE of £3,727 represent the strongest predictive performance in this study, and its stability across training and test sets (overfitting gap of only 0.17) makes it reliable in production. However, organizations choosing Random Forest must be prepared to invest in post-hoc explanation tools such as SHAP values and must accept that the resulting explanations are approximations rather than exact representations of the model's logic.

Never Use Linear Regression for CLV Prediction

Linear Regression is not a viable CLV prediction model in this context. Its test R^2 of 0.3022 compared to a training R^2 of 0.9766 reveals severe overfitting driven by its inability to represent the non-linear, threshold-driven relationships between RFM variables and future customer value. Any organization currently using linear regression as its CLV model is systematically misallocating marketing resources as a result, and migration to either MARS or Random Forest should be treated as a priority.

Do Not Deploy Gradient Boosting Without Extensive Tuning

Gradient Boosting should not be used with default hyperparameter settings under any circumstances. Its test R^2 of -1.4598 in this study worse than predicting the mean value for every customer demonstrates that without careful configuration including learning rate optimization, depth constraints, early stopping, and regularization, the algorithm produces predictions that would actively harm decision-making. If an organization wishes to explore Gradient Boosting for CLV prediction, a rigorous hyperparameter search with cross-validated early stopping is a non-negotiable prerequisite.

Enforce Temporal Validation in All Future Model Development

Regardless of which model is chosen, all CLV models developed within the organization must be evaluated using a strict temporal validation design features computed from historical data only,

targets measured from future data only. Models evaluated without this design produce inflated accuracy estimates that do not hold in production, creating false confidence and eventual failure.

Summary Decision Guide:

Deployment Context	Recommended Model
Regulated environment requiring explainability	MARS
Stakeholder-facing segmentation and strategy	MARS
Fully automated system-to-system prediction	Random Forest
Accurate predictions with interpretability needed	MARS
Linear relationships assumed	Not recommended use MARS or Random Forest
Gradient Boosting with default settings	Not recommended tune extensively first

5.8 Limitations of the Study

Dataset specificity. The findings are grounded in a single retailer's transactions during a specific two-year period. Threshold values, segment compositions, and model accuracy levels may vary across different retail categories, geographies, time periods, and business models. Practitioners applying this methodology to their own data should treat the specific numerical findings as illustrative rather than directly transferable.

Feature set constraints. The restriction to three RFM features, while theoretically justified and practically convenient, leaves potentially informative signals product category preferences,

acquisition channel, geographic clustering, promotional responsiveness unused. Models incorporating richer feature sets would likely achieve higher accuracy and might alter the threshold values identified by MARS.

High-value customer prediction quality. The heteroscedasticity observed in residual plots confirms that all four models perform meaningfully worse for high-value customers than for the customer population as a whole. Because these customers are of the greatest strategic importance, their prediction quality warrants specific attention. Separate models trained exclusively on the high-value segment, or probabilistic approaches that generate prediction intervals rather than point estimates, could substantially improve decision quality for the Champions and At-Risk High-Value segments.

Gradient Boosting configuration. The results for Gradient Boosting reflect default rather than optimized hyperparameter settings. A properly tuned GBM model with grid-searched learning rate, depth, and iteration count, combined with early stopping and regularization might achieve accuracy competitive with or superior to Random Forest. The current results should be interpreted as a benchmark for untuned Gradient Boosting, not as evidence of the algorithm's ceiling performance.

5.9 Summary

The principal empirical findings of this study are:

Non-linear models substantially outperform Linear Regression for CLV prediction (Random Forest test $R^2 = 0.7244$, MARS test $R^2 = 0.6910$ vs. Linear Regression test $R^2 = 0.3022$), confirming that the RFM-to-CLV relationship contains non-linearities that must be modeled explicitly.

MARS and Random Forest perform comparably on the primary accuracy metrics, with a gap of approximately three percentage points in R^2 and £220 in RMSE a difference that is modest relative to the interpretability advantage of MARS.

Gradient Boosting applied with default hyperparameters fails catastrophically (test $R^2 = -1.4598$), demonstrating that algorithmic sophistication is not a substitute for careful

configuration and that default settings can produce predictions that actively mislead decision-making.

MARS identifies behaviorally meaningful threshold values in the RFM features particularly a recency threshold at 12 days and a frequency threshold at 2–3 transactions that translate directly into actionable customer segmentation rules.

The six-segment customer classification scheme derived from MARS outputs spans a CLV range from £616 (Lost Customers) to £10,855 (Champions), a ratio of more than 17:1, underscoring the enormous value of differential marketing strategies calibrated to predicted customer value.

CHAPTER 5

CONCLUSION

This dissertation set out to address four interconnected problems in Customer Lifetime Value prediction: the inadequacy of linear models for a non-linear prediction task, the opacity of the most accurate machine learning alternatives, the prevalence of data leakage in model evaluation designs, and the gap between statistical predictions and actionable business strategy. The empirical work has provided clear answers to all four.

The first conclusion is unambiguous: linear regression is not a viable approach for CLV prediction in retail contexts where behavioral patterns exhibit the threshold effects and interaction dynamics documented in this dataset. The collapse from a training R^2 of 0.977 to a test R^2 of 0.302 is not a marginal performance shortfall it represents a fundamental structural mismatch between the model's assumptions and the data-generating process. Retail analytics practitioners who continue to rely on linear regression for CLV prediction are working with a tool that is inappropriate for the task, and the cost of this methodological conservatism is systematic misallocation of marketing resources.

The second conclusion concerns the accuracy-interpretability trade-off that has dominated much of the applied machine learning debate. The empirical evidence from this study does not support the view that high accuracy and meaningful interpretability are in fundamental conflict. MARS achieves a test R^2 of 0.691 substantially better than linear regression, and only three percentage points below the study's most accurate model while generating predictions that are fully transparent, mathematically explicit, and directly translatable into operational decision rules. The three-point accuracy gap between MARS and Random Forest exists, but in a business context where the ability to explain a model's logic to a regulator, a board, or a marketing manager carries real value, it does not represent a decisive argument for the less interpretable alternative. For most retail CLV applications, MARS provides the better overall solution.

The third conclusion is methodological: the temporal validation design implemented in this study is not a technical nicety but a substantive requirement for reliable CLV model evaluation. Performance estimates generated without strict separation of feature and target windows cannot

be trusted as indicators of real-world model behavior. The frequency with which this requirement is violated in the published literature as documented by Chou et al. (2022) and confirmed by the broader review conducted in Chapter 2 means that many published CLV benchmarks are not reproducible in production settings. This study provides both the methodological template and the empirical demonstration of what rigorous temporal validation looks like in practice.

The fourth conclusion addresses the actionability gap between statistical predictions and business strategy. The six-segment customer classification scheme derived from MARS outputs Champions, Loyal Customers, At-Risk High-Value Customers, New Customers, Lost Customers, and Others demonstrates that the output of a well-designed CLV model need not be an abstract number attached to a customer record. It can be a complete behavioral taxonomy, with each category defined by explicit criteria, characterized by a predicted CLV distribution, and accompanied by specific marketing strategies calibrated to the segment's behavioral profile and strategic potential. The 17:1 ratio in average CLV between Champions and Lost Customers underscores the magnitude of the strategic opportunity that CLV-based segmentation unlocks.

The failure of Gradient Boosting deserves recognition as a finding of independent significance. The tendency to assume that more sophisticated algorithms will yield better results a form of methodological cargo-cult thinking that is widespread in applied analytics is directly contradicted by the results of this study. A model that achieves a test R^2 of negative 1.46 is not merely underperforming; it is actively harmful to the decision-making processes it is supposed to inform. The practical implication is that algorithm selection, configuration, and validation require domain expertise and analytical discipline, not merely access to a machine learning library.

Several ⁸⁷ directions for future research emerge naturally from this work. The incorporation of features beyond the RFM framework including product category preferences, acquisition channel, promotional response history, and customer demographic data where available would likely narrow the accuracy gap between MARS and ensemble methods by providing the former with more informative signals. The development of dedicated prediction models for high-value customer segments addressing the heteroscedasticity limitation identified in the residual analysis could substantially improve the decision quality for Champions and At-Risk High-Value

customers, precisely the segments where prediction accuracy has the highest strategic value. And the extension of the temporal validation framework to longer prediction horizons three years, five years would test the stability of the segment classifications and the durability of the threshold values identified by MARS across different phases of the customer lifecycle.

From the perspective of regulatory and ethical accountability in algorithmic decision-making, the trajectory is clear: transparency requirements will continue to tighten, and the business case for inherently interpretable models will strengthen accordingly. MARS, as a method that achieves competitive accuracy with full structural transparency, is positioned to become increasingly relevant as organizations bring their analytical practices into alignment with emerging accountability standards. This study provides both the empirical foundation and the practical framework to support that alignment.

REFERENCES

- ²⁵ Boehmke, B. C., & Greenwell, B. M. (2019). *Hands-on machine learning with R*. Boca Raton: CRC Press. Retrieved from <https://www.routledge.com>
- ³ Chou, P. C., Chuang, H. H., Chou, Y., & Liang, T. (2022). Predictive analytics for customer repurchase: Integration of buy-till-you-die modeling and machine learning. *European Journal of Operational Research*, 296(2), 635–651. Retrieved from <https://doi.org/10.1016/j.ejor.2021.04.015>
- ¹⁴ Fader, P. S., & Hardie, B. G. (2009). Probability models for customer-base analysis. *Journal of Interactive Marketing*, 23(1), 61–69. Retrieved from <https://doi.org/10.1016/j.intmar.2008.11.003>
- ²⁴ Friedman, J. H. (1991). Multivariate adaptive regression splines. *The Annals of Statistics*, 19(1), 1–67. Retrieved from <https://projecteuclid.org>
- ¹⁸ Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction* (2nd ed.). New York: Springer. Retrieved from <https://link.springer.com>
- ³⁰ Jackson, B. B. (1985, November–December). Build customer relationships that last. *Harvard Business Review*, 63(6), 120–128. Retrieved from <https://hbr.org>
- ⁹⁰ Parry, M. E. (2013). The determination of service quality and its effect on customer lifetime value. *South African Journal of Business Management*, 44, 47–64. Retrieved from <https://journals.co.za>
- ¹² Schmittlein, D. C., Morrison, D. G., & Colombo, R. (1987). Counting your customers: Who are they and what will they do? *Management Science*, 33(1), 1–24. Retrieved from <https://www.jstor.org>

⁷ Sun, Y., Liu, H., & Gao, Y. (2023). Research on customer lifetime value based on machine learning algorithms and customer relationship management analysis model. *Heliyon*, 9(2), e13384. Retrieved from <https://doi.org/10.1016/j.heliyon.2023.e13384>

⁵ Vafeiadis, T., Diamantaras, K. I., Sarigiannidis, G., & Chatzisavvas, K. C. (2015). A comparison of machine learning techniques for customer churn prediction. *Simulation Modelling Practice and Theory*, 55, 1–9. Retrieved from <https://doi.org/10.1016/j.simpat.2015.03.003>

Verhoef, P. C., & Donkers, B. ²³ (2005). The effect of acquisition channels on customer loyalty and cross-buying. *Journal of Interactive Marketing*, 19(2), 31–43. Retrieved from ⁸ <https://doi.org/10.1002/dir.20037>

Wolpert, D. H., & Macready, W. G. (1997). No free lunch theorems for optimization. *IEEE Transactions on Evolutionary Computation*, 1(1), 67–82. Retrieved from <https://doi.org/10.1109/4235.585893>

ANNEXURE

Annexure A : Dataset Variable Descriptions -

Variable	Type	Description
InvoiceNo	Categorical	Unique invoice identifier. Prefix "C" denotes cancellation.
StockCode	Categorical	5-digit ¹³ product identifier.
Description	Text	Product name.
Quantity	Numeric	Units per transaction line. Negative values indicate returns.
InvoiceDate	DateTime	Date and time of transaction.
UnitPrice	Numeric	Per-unit price in GBP.
CustomerID	Categorical	Unique 5-digit customer identifier.
Country	Categorical	Customer country.

Annexure B : Customer Segment Classification Rules –

Champions:

recency_days $\leq$ 12 AND frequency $\geq$ 3 AND monetary $\geq$ £1,071

Loyal Customers:

recency_days $\leq$ 33 AND frequency $\geq$ 2 AND monetary $\geq$ £465

At-Risk High-Value:

$\text{recency_days} \geq 73$ AND $\text{frequency} \geq 3$ AND $\text{monetary} \geq \text{£1,071}$

New Customers:

$\text{recency_days} \leq 12$ AND $\text{frequency} \leq 1$

Lost Customers:

$\text{recency_days} \geq 73$ AND $\text{frequency} \leq 1$

Others:

All remaining customers

Annexure C: Abbreviations -

- ⁸⁰ CLV: Customer Lifetime Value
- RFM: Recency, Frequency, Monetary
- ⁷⁰ RMSE: Root Mean Squared Error
- MAE: Mean Absolute Error
- MARS: Multivariate Adaptive Regression Splines

MRP

ORIGINALITY REPORT

9%

SIMILARITY INDEX

7%

INTERNET SOURCES

6%

PUBLICATIONS

5%

STUDENT PAPERS

PRIMARY SOURCES

1	www.mdpi.com Internet Source	1 %
2	dspace.dtu.ac.in:8080 Internet Source	< 1 %
3	www.irjmets.com Internet Source	< 1 %
4	Dothang Truong. "Demystifying AI – Data Science and Machine Learning Using IBM SPSS Modeler", CRC Press, 2025 Publication	< 1 %
5	www.aasmr.org Internet Source	< 1 %
6	Submitted to City University Student Paper	< 1 %
7	phd.mater.uni-mate.hu Internet Source	< 1 %
8	Submitted to The Scientific & Technological Research Council of Turkey (TUBITAK) Student Paper	< 1 %
9	Submitted to Indira Gandhi Delhi Technical University for Women Student Paper	< 1 %
10	Pradeev B, Kaviya K, Kiruba D, Lakshmipriya M, Loganayagi M. "Development of Fault Tolerant Digital System Using Triple Modular	< 1 %

Redundancy", International Research Journal on Advanced Engineering Hub (IRJAEH), 2026

Publication

11	Submitted to University of Lancaster–Main Account	< 1 %
	Student Paper	
12	www.i-scholar.in	< 1 %
	Internet Source	
13	Submitted to Berlin School of Business and Innovation	< 1 %
	Student Paper	
14	www.mla.vgtu.lt	< 1 %
	Internet Source	
15	ijsrem.com	< 1 %
	Internet Source	
16	erepo.uef.fi	< 1 %
	Internet Source	
17	Submitted to Columbia University	< 1 %
	Student Paper	
18	ttu-ir.tdl.org	< 1 %
	Internet Source	
19	Submitted to Bogazici University	< 1 %
	Student Paper	
20	Submitted to Delhi Metropolitan Education	< 1 %
	Student Paper	
21	Submitted to Islington College,Nepal	< 1 %
	Student Paper	
22	publikationen.bibliothek.kit.edu	< 1 %
	Internet Source	
23	pure.uvt.nl	< 1 %
	Internet Source	

24

www.biorxiv.org

Internet Source

< 1 %

25

www.theseus.fi

Internet Source

< 1 %

26

core.ac.uk

Internet Source

< 1 %

27

sitams.ac.in

Internet Source

< 1 %

28

thecomputerscienceprofessor.com

Internet Source

< 1 %

29

al-kindipublishers.org

Internet Source

< 1 %

30

libstore.ugent.be

Internet Source

< 1 %

31

J. Krishnamoorthi, B. Vinoth Kumar, L. Thulasimani, D. Dhanalakshmi, Rajiv Gupta. "Advancing Societally Relevant Applications of Knowledge through Scientific Research", CRC Press, 2026

Publication

< 1 %

32

Udara Yedukondalu, V Vijayasri Bolisetty. "Advancing Innovation through AI and Machine Learning Algorithms – Computational Intelligence for Virtual System Optimization. A proceeding of ICMVRCET – 2025", CRC Press, 2025

Publication

< 1 %

33

Submitted to Dublin Business School

Student Paper

< 1 %

34

Submitted to INTI Universal Holdings SDM BHD

Student Paper

< 1 %

35	Submitted to Queensland University of Technology	< 1 %
	Student Paper	
36	Submitted to Yoobee	< 1 %
	Student Paper	
37	hal.inria.fr	< 1 %
	Internet Source	
38	jurnal.itscience.org	< 1 %
	Internet Source	
39	123docz.net	< 1 %
	Internet Source	
40	Submitted to Manipal University	< 1 %
	Student Paper	
41	Submitted to The University of the West of Scotland	< 1 %
	Student Paper	
42	Yinglin Xia, Jun Sun. "Machine Learning for Microbiome Statistics", CRC Press, 2026	< 1 %
	Publication	
43	kubanni-backend.abu.edu.ng	< 1 %
	Internet Source	
44	Submitted to Brunel University	< 1 %
	Student Paper	
45	Submitted to Coventry University	< 1 %
	Student Paper	
46	www.anowa.reapress.com	< 1 %
	Internet Source	
47	www.edocs.maseno.ac.ke	< 1 %
	Internet Source	
48	Submitted to Aston University	< 1 %
	Student Paper	

49	Submitted to Florida State University Student Paper	< 1 %
50	Submitted to University of Central Lancashire Student Paper	< 1 %
51	e-research.siam.edu Internet Source	< 1 %
52	research.usq.edu.au Internet Source	< 1 %
53	Shourabizadeh, Hamed. "Advanced Machine Learning Techniques for Survival Analysis in Medical Data", University of Toronto (Canada) Publication	< 1 %
54	diglib.tugraz.at Internet Source	< 1 %
55	Submitted to Häme University of Applied Sciences Student Paper	< 1 %
56	Submitted to University of Wollongong Student Paper	< 1 %
57	cardio.jmir.org Internet Source	< 1 %
58	ijeais.org Internet Source	< 1 %
59	jlupub.ub.uni-giessen.de Internet Source	< 1 %
60	primerascientific.com Internet Source	< 1 %
61	thesis.eur.nl Internet Source	< 1 %
62	tohoku.repo.nii.ac.jp Internet Source	< 1 %

63	wiredspace.wits.ac.za Internet Source	< 1 %
64	Submitted to National Institute of Technology, Warangal, Telangana, India Student Paper	< 1 %
65	Submitted to Panjab University Student Paper	< 1 %
66	Submitted to University of Ulster Student Paper	< 1 %
67	datacalculus.com Internet Source	< 1 %
68	www.desjardins.com Internet Source	< 1 %
69	www.dspace.dtu.ac.in:8080 Internet Source	< 1 %
70	www.jetir.org Internet Source	< 1 %
71	Ankur Awadhiya. "Fundamentals of Machine Learning for Life Sciences – A Simple, Step-by-Step Approach with Examples in R", CRC Press, 2026 Publication	< 1 %
72	Cenk Temizel, Salih Tutun, Celal Hakan Canbaz, Ekrem Alagoz et al. "Artificial Intelligence in the Energy Industry – Theory, Case Studies, and Applications", CRC Press, 2026 Publication	< 1 %
73	European Journal of Marketing, Volume 47, Issue 10 (2013–09–28) Publication	< 1 %

74 Gygi, Jeremy. "Supervised Integration of Multi-Omics Immune Profiles via Latent Factor Modeling.", Yale University $< 1\%$
Publication

75 Siddhartha Roy, Soumya Sen, Agostino Cortesi. "Intelligent Systems – Emerging Trends and Challenges in Deep Neuro-Fuzzy Models and Explainable Artificial Intelligence", CRC Press, 2026 $< 1\%$
Publication

76 Webb, . "Rule and Decision Tree Induction", Statistical Pattern Recognition $< 1\%$
Webb/Statistical Pattern Recognition, 2011.
Publication

77 aaairj.com $< 1\%$
Internet Source

78 arxiv.org $< 1\%$
Internet Source

79 dspace.daffodilvarsity.edu.bd:8080 $< 1\%$
Internet Source

80 jaguar.uma.pt $< 1\%$
Internet Source

81 jurnal.uts.ac.id $< 1\%$
Internet Source

82 link.springer.com $< 1\%$
Internet Source

83 marjoleinf.r-universe.dev $< 1\%$
Internet Source

84 opus4.kobv.de $< 1\%$
Internet Source

85 vdoc.pub $< 1\%$
Internet Source

86

www.econ.uzh.ch

Internet Source

< 1 %

87

www.ijsred.com

Internet Source

< 1 %

88

Albert Wong, Andres Vilorio Garcia, Yew-Wei Lim. "A data-driven approach to customer lifetime value prediction using probability and machine learning models", *Decision Analytics Journal*, 2025

Publication

< 1 %

89

Khera, Shikha N., and Karishma Gulati. "Reward system and TKSb of employees of IT organisations: an empirical study", *International Journal of Knowledge Management Studies*, 2015.

Publication

< 1 %

90

Dipak Jain, Siddhartha S. Singh. "Customer lifetime value research in marketing: A review and future directions", *Journal of Interactive Marketing*, 2002

Publication

< 1 %

91

Ozgur Kisi. "Modeling reference evapotranspiration using three different heuristic regression approaches", *Agricultural Water Management*, 2016

Publication

< 1 %

92

Tayyeba Bashir, Tan Zhongfu, Burhan Sadiq, Ammara Naseem. "How AI competencies can make B2B marketing smarter: strategies to boost customer lifetime value", *Frontiers in Artificial Intelligence*, 2024

Publication

< 1 %

Exclude quotes	Off	Exclude matches	Off
Exclude bibliography	Off		