International Journal of Scientific and Management Research
Volume 9 Issue 3 (March) 2026
ISSN: 2581-6888
Page: 1-19

Predicting Customer Lifetime Value to Inform Product Investment Decision

Ayokunmi Sodamola[1]*, Emmanuella Wiafe[2], Chukwuka Stanley Ekeocha[3], Rianat Abbas[4], & Mohamed Sheriff Jalloh[5]
[1]Department of Management Information Systems, Baylor University, Texas, **United States**
[2]Department of Biotechnology, Northeastern University, Boston, **United States**
[3]Department of Business Administration, Indiana University, Indiana, **United States**
[4]Department of Management Information Systems, Baylor University, Texas, **United States**
[5]Department of Business Administration, Westcliff University, California, **United States**
DOI - http://doi.org/10.37502/IJSMR.2026.9301

Abstract

The growing intensity of retail market competition has compelled organizations to move beyond transactional performance metrics toward forward-looking measures of customer economic value. Customer Lifetime Value (CLV) has emerged as a strategically significant construct for understanding the long-term revenue potential of individual customer relationships, yet its systematic application to product investment decision-making remains methodologically underdeveloped in the literature. This study addresses that gap by employing ensemble machine learning models to predict CLV tiers and map the resulting classifications onto a structured product investment framework. Utilizing a publicly available retail dataset of 736 customers sourced from Kaggle, CLV tiers were engineered as a composite of tenure-adjusted spend, average order value, and retention probability, and customers were classified into Low, Medium, and High value segments. Two ensemble classifiers, Random Forest and XGBoost were trained, tuned using five-fold cross-validation with grid search, and evaluated on a 30% holdout test set. Random Forest achieved superior overall performance with an accuracy of 63% and a weighted F1 score of .62, outperforming XGBoost across all evaluation metrics. Both models consistently identified average order value and tenure months as the most dominant predictors of CLV tier, confirming that spending intensity and relationship longevity are the primary drivers of long-term customer value. CLV-to-product investment mapping revealed that Home and Garden and Groceries categories attract the highest concentration of high-value customers, while Electronics, despite lower penetration, generates the highest average spend among high-value customers. These findings demonstrate that structured CLV assessment, operationalized through ensemble machine learning, provides organizations with a robust and evidence-based foundation for aligning product investment priorities with the customers most likely to generate sustainable long-term revenue growth.

Keywords: Customer Lifetime Value, Ensemble Machine Learning, Product Investment Decision, Random Forest, XGBoost

1. Introduction

The business world in the global market has become significantly competitive, which has forced organizations to fundamentally re-evaluate value creation and sustainability (Ali &

Shabn, 2024). Growing market saturation, the narrowing of product differentiation margins, and a catastrophic rise in the cost of acquiring customers, 222 percent in the past eight years, have rendered it increasingly unsustainable to support the volume and product-oriented approach to the market exclusively (Wong et al., 2025). Formal data strategies have been embraced by 68 percent of organizations and 89 percent of executives intend to increase spend on analytics Hydrogen BI, marking a tipping point to intelligence-based business models that are customer-focused (Segarra-Moliner & Bel-Oms, 2023). The comprehension of the personal economic value of a client has ceased being a marketing luxury and become a main strategic requirement that determines how companies set priorities on their resources, relationships, and developmental growth paths (Oghenemaro, 2025).

At the center of this customer-focused transformation is the Customer Lifetime Value (CLV) construct or measure that approximates how much net profit an organization is likely to make out of a customer over the course of their relationship or lifetime (Nyakeri, 2025). As opposed to transactional metrics which measure performance at a specific point in time, CLV follows a longitudinal perspective by combining purchase frequency, average order value, retention probability, and discount rates as one future-looking estimation (Ekinci et al., 2014). Wong et al. (2025) and Cowan et al. (2023) indicated that one out of five customers bring 80% of the revenue of a company, which makes the contribution of customers uneven indeed. CLV provides organizations with the analytical language to determine, measure, and act on this disparity and therefore, it is one of the most strategically pertinent measures that contemporary business can have in their pursuit of sustainable competitive advantage (Al Rafi and Yassar, 2025).

In addition to the descriptive role, CLV has been used more as a decision-making tool in distributing marketing and operational resources. When organizations use CLV to priorities resource allocation, they stand in better positions to invest in the relationships with high-value customers other than uniformly distributing their efforts (Nyakeri, 2025). Bain and Company have documented that a 5% customer retention increase will lead to a 25 to 95 percent increase in profit which has motivated many companies to make their retention investment decision based on CLV (Oghenemaro, 2025). Rising to the recent economic challenges, 67 percent of brands have pivoted away on the acquisition aspect towards the retention aspect Shopify, a more generalized acknowledgement that CLV-informed allocation will yield better returns (Awaad et al., 2024; You, 2025). This re-orientation leads inevitably to the question whether it is possible to apply CLV intelligence, which is being used in the field of customer management, to the field of investing in products.

Product investment decisions - including the creation, improvement and prioritization of product features, lines, portfolios are among the most significant and resources intensive decisions that organizations make (Ali & Shabn, 2024). Wong et al. (2025) indicated that 79 percent of the executives consider product management to be critical to the success of their company, but only 12 percent of the businesses have a fully developed product management process. This disjunction indicates a long-standing problem: investment decisions are often made based on short-term revenue indicators, competitive pressures or internal lobbying instead of doing proper, customer-based analysis (Nyakeri, 2025). Product teams guided by data are 2.9 times more likely to have products that achieve their business objectives, but the processes that guide data-driven methods do not tend to systematically utilize CLV as an input

and thus are not using all the analytical capabilities possible (Cowan, Mercury, & Khraishi, 2023).

Although the CLV models have become more sophisticated and their strategic importance has become widely recognized, a major gap still exists between CLV intelligence and product investment practice. Organizations that calculate CLV usually use it when considering marketing, and retention functions, and product investment decisions remain steered by distinguishable systems - commonly on the basis of market size, feature demand, or competitive benchmarking (Ekinci, Uray, & Ulengin, 2014). It is no surprise that the CLV is a measurement that is not appropriately measured by many companies, with only half calculating the CLV-to-CAC ratio, therefore it is unlikely to find its way into the decision-making about product portfolio (Al Rafi & Yassar, 2025). It is such soloing of customer value intelligence of product strategy that is a lapsed opportunity of a significance scale as companies are increasingly pressured to demonstrate long-term, customer-based investments with long-term customer evidence (Segarra-Moliner and Bel-Oms, 2023).

Attempts to establish customer value analytics and product strategy are not completely unmentioned in the literature. Academics have discussed the issue of CLV segmentation as a foundation of personalized product positioning, and practitioners have experimented with the use of cohort analysis to guide feature prioritization. High-tech CLV methods, such as machine learning algorithms, have been shown to be able to process big volumes of customer data, determine non-linear relationships, and make predictions that are highly accurate which in the context of product investment promises a definite opportunity (Oghenemaro, 2025). Nonetheless, these developments largely have been in the sphere of customer retention and marketing optimization. Their conversion into operationalizable and structured frameworks of product investment decision-making has not been carried out in sufficient depth and breadth across industry settings and thus leave a gap in the scholarly and practical literature that the current study aims to fill.

It is a joint coming together of enriched customer information, more potent predictive modelling frameworks, and increasing organizational interest in evidence-based choices to make investment, that makes an integrated CLV assessment framework, which explicitly deals with and influences product investment selections, a call to action. The analytics-driven firms are 23 times more likely to win customers, 6 times more likely to keep customers and 19 times more likely to be profitable, a performance margin that highlights the potential of bridging the analytics gap between customer value and product strategy (Cowan, et al., 2023). Through systematically evaluating CLV per customer segment, and overlaying the evaluations on product investment priority, organizations can tune their development pipelines to the customers that are likely to bring net present value to them over time. The paper will serve that quest by advancing and discussing a systematic model in terms of which CLV measure will help to make product investments in a meaningful and reliable way.

2. Literature Review

2.1 Conceptual Foundations of Customer Lifetime Value

Customer Lifetime Value as a concept is based intellectually on the direct marketing literature of the 1980s, in which the identified users of the concept first realized that the long-term potential of a customer relationship was a more useful performance metric than the individual

transaction. The initial formulations were quite primitive and viewed CLV as the mere extrapolation of previous purchasing behavior into the future. Instead, it was only in the 1990s that CLV was more thoroughly theoretically grounded, primarily through the efforts of Blattberg and Deighton (1996) who believed that companies must invest in customers until the marginal cost of retention was equal to the marginal return - a concept that, once again, reinstated CLV as a resource allocation device, instead of a reporting instrument.

Fundamentally, CLV is the present value of all cash flows in the future that can be accredited to a customer relationship and that are discounted to consider the time value of money and the risk of customer losses. It is formally combined of three main forces; the amount of margin contributed by each transaction, the repetition and persistence of purchases over time, and the discount rate applied to future cash flows. This definition contrasts CLV with retrospective measures like customer revenue or historical expenditure, which instead premises it on probabilistic forecasting and long-term financial reasoning. Kumar and Shah (2009) also added a theoretical construct to the fact that CLV must not be thought of in a vacuum but rather as a component of a larger customer equity model, whereby the total lifetime value of the entire customer base of a company is a quantifiable and manageable strategic asset.

The development of CLV theory has followed the improvement of the availability of data and computing power closely. Out of plain recency-frequency-monetary scoring during the 1990s, probabilistic hazard models during the 2000s, machine learning architectures during the 2010s, and others, the conceptual frameworks of CLV have increasingly allowed greater complexity, heterogeneity, and predictive ambitions. The consistent aspect of these cycles has been the underlying assumption, which is that it is not known how much economic value one relationship with a customer will have in the future, and that knowledge would play a central role in making intelligent business decisions.

2.2 Customer Lifetime Value Measurement Approaches

The Customer Lifetime Value metric has changed significantly in the past decades and resulted in a variety of methodological designs with different levels of complexity, data needs, and prediction capabilities. All these methods fall broadly into four groups: heuristic scoring schemes, probabilistic statistical schemes, econometric regression schemes, and machine learning schemes, each of which describes the analytical ability and data conditions of the respective time, and each has its own benefits and constraints available to the practitioner.

Recency, Frequency, and Monetary (RFM) model is the oldest heuristic CLV approximation model with the greatest adoption. RMF came into existence in direct mail marketing where customers are categorized in terms of the time of the last purchase, the frequency of purchase and the amount of the purchase. It has maintained popularity because of its simplicity and interpretability, that is, it does not require any complex infrastructural support in statistics and produces actionable customer segments without resorting to complex modelling. Nevertheless, RFM is essentially descriptive and not predictive; it generalizes previous behavior without making a formal prediction of the likelihood of future transaction or considering the time value of money. Consequently, RFM is also handy in operations-level segmentation, but structurally limited in producing valid CLV estimates, especially in non-contractual customer interactions where the time of purchase is stochastic.

Schmittlein, Morrison and Colombo (1987) Pareto/NBD model was a breakthrough concept in methodology with a probabilistic approach to customer buying and consumer dropout behavior. This model assumes that the customers buy at a single rate of Poisson when active and defecting at the unobservable point which is an exponential rate, and both rates are non-homogeneous among the customer population governed by gamma distributions. This two-process specification allows the model to approximate the probability of a customer remaining active and the number of transactions that the customer will have in the future, giving a theoretically justified CLV estimate. This model was later refined by the BG/NBD model of Fader, Hardie and Lee (2005) which substituted the exponential dropout assumption with a beta-geometric model, which is more convenient to work with and statistically more accurate when large proportions of one-time buyers or high dropout rates exist in a dataset. The two models have been shown to perform quite well in non-contractual environments and continue to be common benchmarks in the academic and practical research of CLV.

Of more recent interest, machine learning methods, such as gradient boosting, random forests, neural networks, as well as survival models have broadened the methodological frontier significantly. These methods are especially adapted to high-dimensional customer data sets where parametric assumptions are unlikely to be true and allow the analyst to capture non-linear relationships, effects of interaction and dynamics that are not well-captured by probabilistic models. Empirical comparisons have usually established that machine learning models perform better than classical approaches on crude predictive accuracy especially when abundant behavioral, contextual, and transactional features exist. Nevertheless, their comparative opaqueness makes them difficult to interpret and use in management, and there is increasing interest in explainable AI methods, which are both predictive and transparent. These four methodological traditions can be used as a complementary set of tools to measure the CLV, and the most suitable one should be selected based on the availability of data, the specifics of the organization, and the target use of the estimated results.

2.3 Customer Lifetime Value and Segmentation Strategies

The customer segmentation is not new to the marketing strategy since it has allowed organizations to get out of the undifferentiated mass marketing strategies and instead focus resources on specific segments with homogeneous features and behaviors. The adoption of CLV in segmentation practice is a major advancement of the field, which will turn the foundation of customer segmentation of demographic or behavioral characteristics to the perspective of the economic worth in future. Instead of segmenting the customers based on their identity or what they already bought, CLV-based segmentation organizes them based on their worth, and, more importantly, their potential worth in the course of their relationship with the company.

The most used CLV segmentation model subdivides customers into value segments, commonly high, medium and low, by their estimated lifetime value scores. This tiering model allows organizations to scale retention investment, service, and product access and the intensity of communications directly proportionate to customer value so that the economically most important relationships would be allocated disproportionate strategic focus. Empirically, it was found by Reinartz and Kumar (2003) that customer lifetime value-based segmentation yields considerably greater marketing returns on investment than tenure or frequency-based proxies of customer profitability, which are systematically concealed by the former.

In addition to binary or ordinal tiering, more elaborate frameworks of segmentation integrate CLV with behavioral and attitudinal components to create a more detailed customer image. Rust, Lemon, and Zeithaml (2004) suggested that customers could be broken down by the value they are currently giving to the firm and how responsive they are to the actions of the firm, in effect, providing an action map of where the incremental value could be achieved through retention and development spending. This dynamic segmentation has become quite popular in data rich markets like financial services, telecommunications and e-commerce where finer behavioral data is collected and used to constantly adjust segment limits and value approximations. Segmentation by CLV is therefore a type of analysis classification and not just the effective allocation of value through nominal operation architecture of the customer portfolio.

2.4 Customer Lifetime Value in Strategic Decision-Making and Resource Allocation

Applicability of the Customer Lifetime Value has increasingly or is increasingly going beyond its marketing analytical origins to emerge as a tool of broader strategic decision-making. As organizations have become more complex in their utilization of customer information, CLV has been used as a steering input in such functions as pricing, channel management, service design and corporate valuation. This functional growth is indicative of an increasing appreciation of the fact that the decisions made in any single part of the organization have an eventual implication on the quality and durability of customer relations - and that CLV offers a shared financial language in which such impacts can be quantified, compared and optimized.

CLV has been valuable especially in the resource allocation realm where it is a tool that can solve competing investment priorities in cases of budgetary constraint. The classical allocation models: allocating resources uniformly to the customer segments or focusing expenditure on the biggest customers often have suboptimal results due to the confusion of the size with the value. This is rectified by CLV-informed allocation, which allocates investment to the customers of whom the future contribution projections are worth the cost of retention and development. Gupta and Lehmann (2005) proved that customer base of the firm which is measured in terms of aggregated CLV is a measurable financial asset - one which can be approached with the same degree of professionalism as capital equipment or intellectual property. This framing brought CLV to a strategic balance sheet consideration, and had direct implications on budgeting, the communication of the implications to investors, and corporate strategy.

At organizational level, CLV has also provided insights into customer acquisition levels, sales force allocation, customer loyalty program and pricing tier architecture. The companies that incorporate CLV into their strategy planning cycles are more likely to separate between the customers that create a short-term revenue stream and those that generate a long-term enterprise value, the fact that becomes especially material in the economic pressure times when trade-offs between acquisition and retention expenditures are most pronounced. Strategic implementation of CLV is, therefore, a novel transition between the reactive management of customers and the value-based proactive portfolio stewards.

2.5 Product Investment Decision Frameworks

The product investment decisions take a center stage in the corporate strategy, which includes the decisions of allocating the financial, human, and technological resources in product

development initiatives, addition of features, expansion of the portfolio and entry into the market. Considering the magnitude of the commitment that is typically involved in such decisions and the unpredictability of the results, there has been a significant academic output on the subject that attempts to offer systematic structures by which organizations can evaluate, prioritize, and implement product investments in a more rigorous and consistent manner.

One of the oldest models of this tradition is the Stage-Gate model, developed by Cooper (1990), which divides the process of product development into several phases, each of which is followed by an evaluative checkpoint, or a gate, against which the decisions to further invest in the project are made, depending on the predefined criteria. The Stage-Gate strategy introduced rigor and responsibility to product investment, institutionalizing go/no-go decision points, minimizing the occurrence of resource investment in ill-conceived projects. It has been observed by critics, though, that the model is largely process-oriented making it little applicable to how to evaluate the customer value implications of competing investment options and its sequential structure does not always suit the iterative and rapid paced product environments that characterize modern digital and technology businesses.

The portfolio management theory has provided a parallel direction to this by utilizing financial portfolio logic to consider product investment as a counterbalancing exercise between risk and result as well as strategic fit. Maps like the McKinsey-GE Matrix and the Boston Consulting Group growth-share matrix have been popularly used to justify product portfolios to map offerings by attractive market and competitive position. Although these tools offer a valuable strategic orientation, they do so at a level of abstraction that ignores customer value dynamics and views market opportunity as an external environmental fact and not a role of the particular customer relationships that a firm has.

In more recent times, agile and lean product developmental strategies have brought in customer feedback loops as an iterative investment validation mechanism, focusing on rapid experimentation over a detailed upfront plan. However, even in the contexts of such systems, the introduction of CLV as a formal input into investment prioritization is still absent to a significant extent. The customer insight has a tendency of being channeled into the process in the form of qualitative channels such as user research, satisfaction surveys and net promoter score but not the type of forward-looking value measurement that CLV gives. This methodological gap directly leads to the fact that there is a need to have a more deliberate and systematic relationship between customer lifetime value intelligence and product investment decision-making.

2.6 The Intersection of Customer Lifetime Value and Product Investment

Despite the similar developments of CLV modelling and product investment models, the academic evidence at the crossroads is quite limited. Very little literature has explicitly investigated how the estimates of CLV can be converted into practical product investment criteria and minimal has also suggested coherent structures to operationalize this relationship in a manner that is replicable and organization-free. Where studied, the relationship has been found to arise as a side observation to more generalized customer equity or marketing strategy research, as opposed to the most rigorous study.

The number of gaps can be identified. To begin with, current CLV models are massively geared towards customer-level decisions, retention, reactivation, and acquisition trade-offs over

product-level investment prioritization. Second, product investment models, although methodologically advanced, invariably consider customer value as an input which comes because of market research but not quantitative lifetime value modelling. Third, the feedback loop between the results of product investment and the CLV trajectory i.e. the product decisions change the retention rates, the rate of purchase and the lifetime value is not specified in theory and is not empirically investigated. All these convergent lacks imply that the introduction of CLV evaluation in product investment decision making is not a mere incremental extension of current practice but a genuinely new analytical and strategic input - which is also sought after in this paper.

3. Methodology

This part outlines the approaches to Customer Lifetime Value (CLV) measurement, and how its systematic estimation can be used in the product investment decision-making process in the retail scenario.

3.1 Research Design

The research design applied in this study is a quantitative and cross-sectional study based on the philosophy of positivism. In this way, it is possible to operationalize CLV as a measurable construct and conduct a regular analysis of the relationships between customer demographic, behavioral, and transactional variables. Ensemble machine learning models are trained and examined in forecasting CLV scores and results are then utilized to come up with an outline of decision making with customer value estimates and product investment priorities.

3.2 Data Source and Sample Characteristics

The data set was sourced out of Kaggle and contains 736 distinct retail customer transactions that include variables such as age, gender, region, channel of acquisition, number of transactions, total spend, average order value, recency, tenure, product category and a binary churn variable. It has five product lines, which include Electronics, Groceries, Clothing, Home and Garden, and Beauty and five acquisition channels. The total churn rate stands at 12.8 and the average customer tenure is 23.6 months. One-hot encoding was applied to categorical variables and continuous variables were normalized using min-max scaling then the dataset was split into a 70% training and 30% holdout test set to evaluate the model.

3.3 CLV Target Variable Construction

Considering that the dataset, there is no pre-computed CLV field, the target variable was engineered with a composite scoring method that combines three fields, tenure-adjusted total spends, average order value, and an estimated retention probability based on retention status and churn status. These elements were aggregated on an equal-weight basis with the resulting scores standardized to a continuous range and the scores served as the dependent variable in both ensemble models.

3.4 Ensemble Machine Learning Models

There are two ensemble models that are used. Breiman (2001) invented the name Random Forest, a bagging technique, which builds numerous decision trees on bootstrapped subsamples and combines the results by averaging. It can deal with feature types of mixtures and stands to generate interpretable feature importance estimates. A boosting-based algorithm, XGBoost

(Chen and Guestrin, 2016) constructs trees one after another, by adjusting the residual errors using gradient descent optimization. XGBoost has been chosen due to its increased empirical performance in structured tabular inputs, CPU performance, and missing values by default. These two models are methodologically complementary, with Random Forest minimizing the variance by building it in parallel and the XGBoost minimizing neither bias nor variance by refining the model. The tuning of the hyperparameters of both models was done through fivefold cross-validation through the grid search, and evaluated based on RMSE, MAE, and R2.

3.5 Ethical Considerations

Since the dataset is publicly accessible through Kaggle and has no personally identifiable information then there is no need to run the research through formal ethics review. All records of customers are subjected to anonymized codes, and the analytical results are reported in the segments or aggregate form. Only data pertaining to use in this research was accessed and processed. The assumptions used in the formation of the CLV variables are all clearly recorded to allow them to be replicated and independently evaluated.

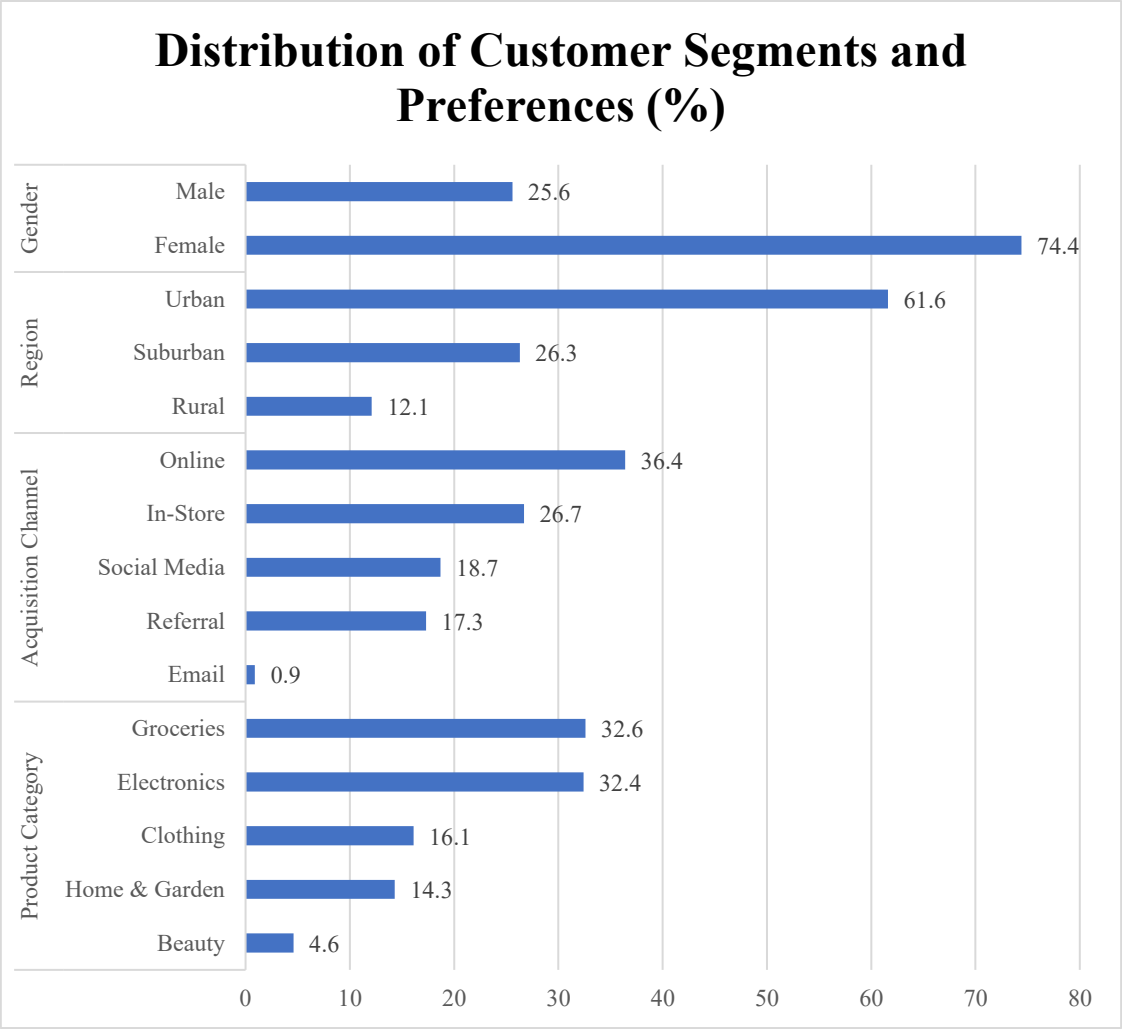


Figure 1: customer demographics and market distribution in percentage

The findings show that most of the respondents were females (74.4%) and urban inhabitants (61.6%) and suburban (26.3) and rural (12.1) were less represented. Online channels had the

greatest number of acquisitions (36.4%), then in-store (26.7) and least in email (0.9%). The largest proportion of products was dominated by groceries (32.6%) and electronics (32.4%) then followed by the smaller proportion of clothing (16.1%), home and garden (14.3%), and beauty (4.6%). This allocation brings out city, female consumer and digitally based consumer trends.

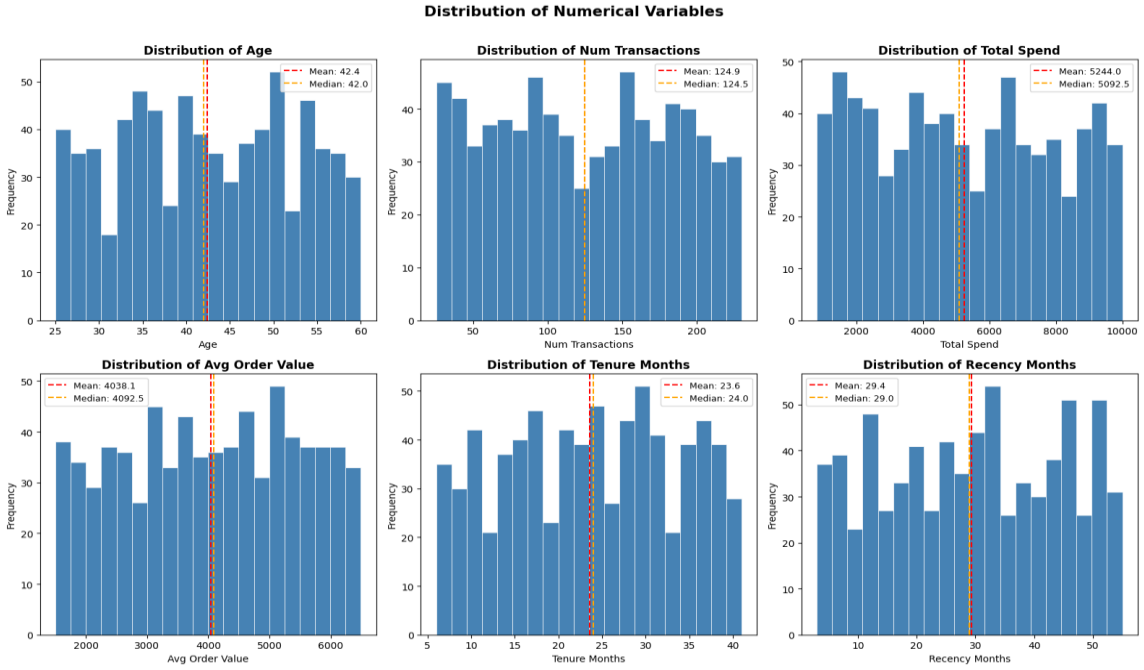


Figure 2: distribution of numerical variables

The analysis of the numerical variables in Figure 2 showed that in the majority of features, there were homogeneous tendencies. The age distribution was relatively equal between 25 and 60 years (M = 42.4, Mdn = 42.0), and num transactions between 25 and 230 (M = 124.9, Mdn = 124.5) indicating little skewness. Total spend (M = 5,244.0, Mdn = 5,092.5) and average order value (M = 4,038.1, Mdn = 4,092.5) had comparatively flat distributions over the range of their values. The tenure months (M = 23.6, Mdn = 24.0) and recency months (M = 29.4, Mdn = 29.0) were also very homogenous and none of the variables showed any severe outliers or any distributional anomalies, thus making the dataset to be suitable to use in ensemble modelling without transformation.

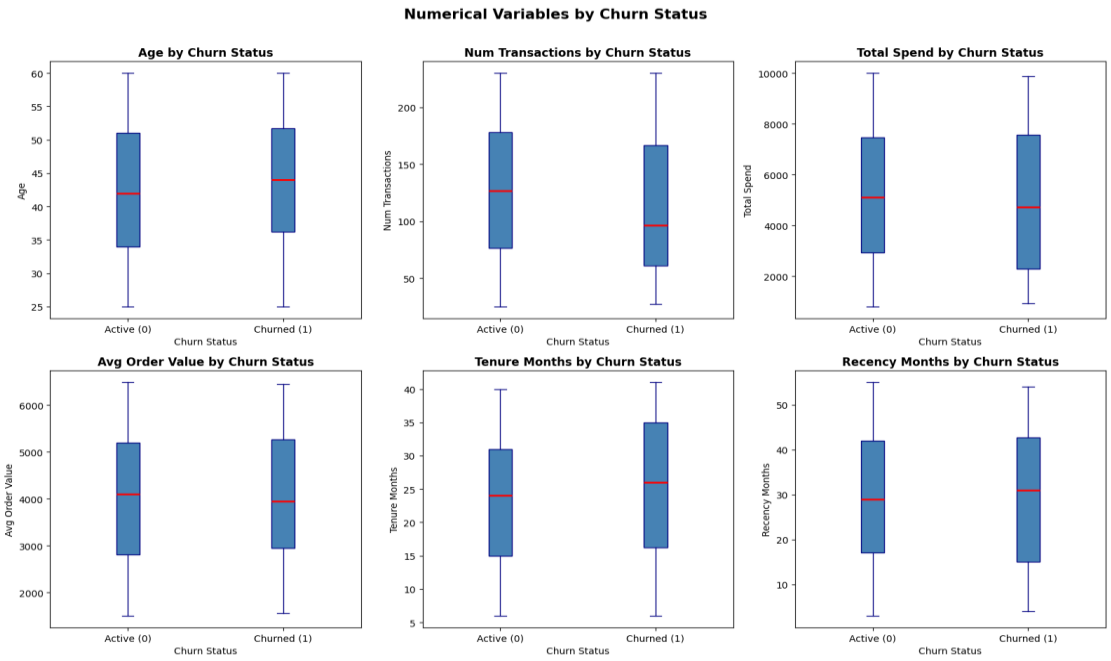


Figure 3: numerical variables by churn rate

The numerical variables versus churn status bivariate analysis as shown in Figure 3 indicated that there were typically small distributional differences between the active and churned customers. The median age of the active customers (Mdn = 42) was lower than the median age of the churned customers (Mdn = 44), but there was a slight difference in the median age of the churn groups. The median number of transactions of churned customers (Mdn = 95) was also significantly lower compared to active customers (Mdn = 125), which means that less frequent purchasers are more vulnerable to lapse. On the same note, the median total spend and median order value were lower with churned customers, and shorter tenure and median recency months were registered by the active customers. In general, the distributions between both sets were quite overlapping, as the weak correlations in Figure 2 imply that no particular numerical variable in isolation can be used to distinguish between active and churned customers.

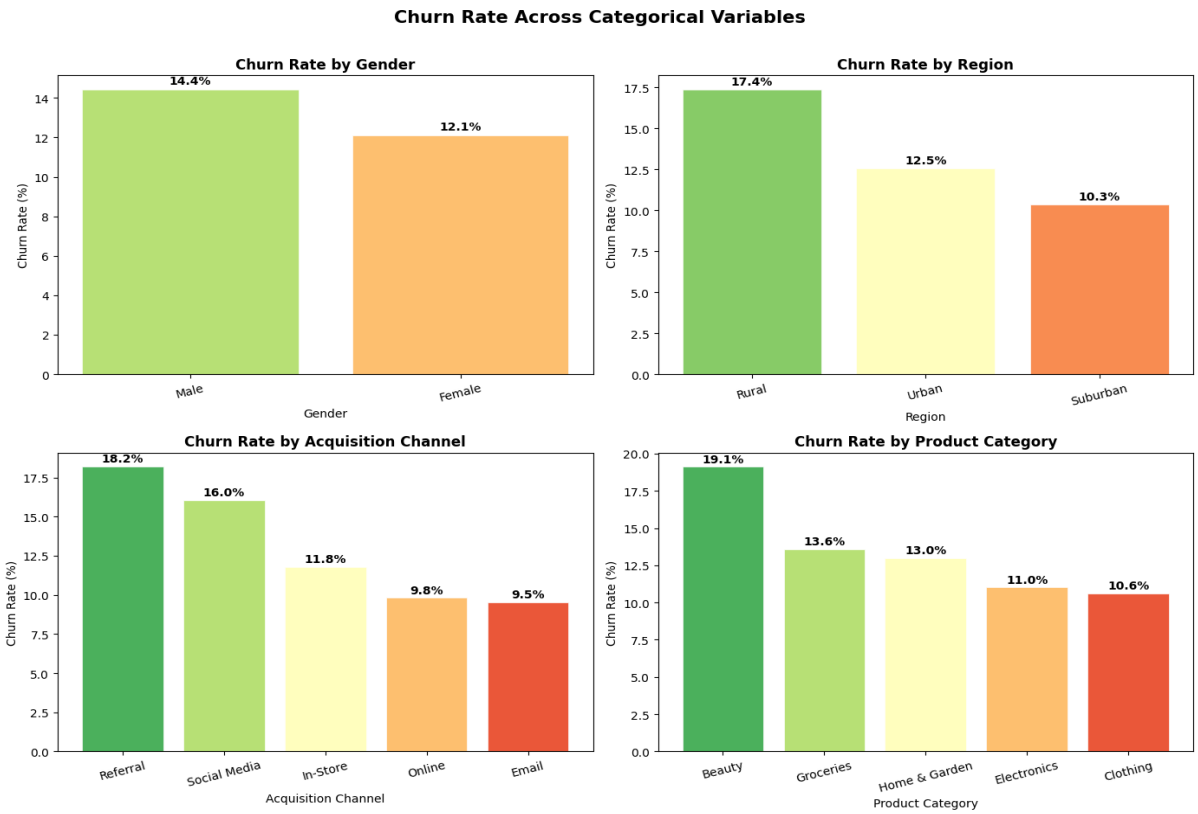


Figure 4: churn rate by gender, region, acquisition channel, and product category

Figure 4 shows the churn rates as per the four categorical variables where there is significant difference in the customer attrition rates across the demographic and behavioral groups. In terms of gender male customers had a higher churn rate (14.4) than female customers (12.1). Rural customers showed the highest churn rate (17.4%), then urban (12.5%) and suburban customers (10.3) so it can be supposed that geographical location can be of significant importance in the retention of customers. The acquisition channel analysis also showed that customers acquired through referrals (18.2%), social media (16.0%), in store (11.8%), online (9.8%), and email (9.5%) channels exited the business the most, showing that the organically acquired customers on the email and online channels are superior to retain. Regarding product category, Beauty had the highest churn rate (19.1%), then the Groceries (13.6%), Home and Garden (13.0%), Electronics (11.0%), and Clothing (10.6%), indicating that those customers who buy the Beauty products were the most at-risk group in the entire data set.

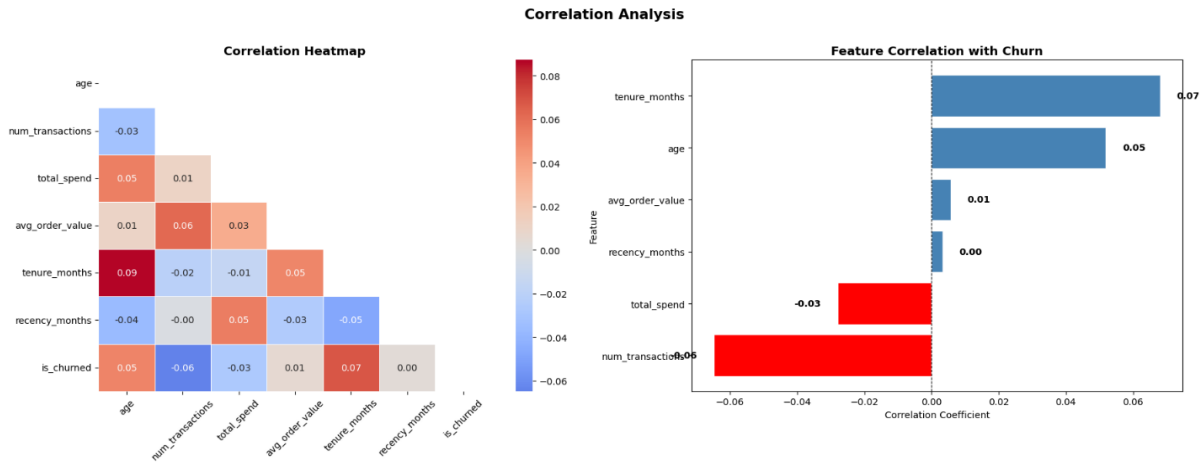


Figure 5: correlation analysis

The correlation analysis showed that all the numerical variables had homogeneous weak correlations. Tenure months showed the most positive relationship with churn (r =.07), then age (r =.05), whereas num-transactions showed the most negative relationship (r = -.06), indicating that the more transactions a customer makes, the less likely he/she will churn. There was a minor negative correlation between total spend (r = -.03) and recency months (r = .00) and a near zero relationship between churn and recency months (r =.00). The predictor variables did not exhibit any multicollinearity concerns, which provided the support of the collective inclusion of the predictors in the ensemble models.

4.3 Hyperparameter Tuning of the Models

A hyperparameter tuning process was conducted for both ensemble models employed in this study. A grid search method was used in conjunction with 5-fold cross-validation, systematically evaluating all possible combinations of predefined parameter values across the training set. The primary objective of the tuning procedure was to identify the parameter combination for each algorithm that maximized the weighted F1 score, a metric prioritized on account of its capacity to balance precision and recall simultaneously across all three CLV tier classes.

Table 1: Optimal Hyperparameters for Random Forest and XGBoost

Model	Hyperparameter	Optimal Value
Random Forest	N_estimators	100
	Max_depth	10
	Min_samples_split	10
	Min_sample_leaf	4
XGBoost	n_estimators	200
	max_depth	3
	learning_rate	0.01
	subsample	0.8
	colsample_bytree	1.0

The optimal Random Forest model comprises an ensemble of 100 decision trees (n_estimators = 100). The parameter max_depth = 10 indicates that each tree was permitted to grow to its full depth without restriction, suggesting that the model required complex, deeply branched trees

to adequately capture the non-linear relationships among customer demographic and behavioral features driving CLV tier classification. The min_samples_split = 10 and min_samples_leaf = 4 settings allow the trees to make fine-grained splits down to individual observations, maximizing the model's capacity to learn subtle patterns within the training data.

The optimal XGBoost model also comprises 200 sequentially constructed decision trees (n_estimators = 200). In contrast to Random Forest, these trees are deliberately shallow, with max_depth restricted to 3 levels, a characteristic feature of gradient boosting that prevents any single tree from overfitting to the training data. The learning rate of 0.01 is notably conservative, controlling the contribution of each successive tree and compelling the model to learn gradually and cautiously across the full boosting sequence, which typically yields stronger generalization on unseen data. The subsample and colsample_bytree parameters, both set to 0.8 and 1.0 respectively, introduce stochastic sampling of observations and features at each boosting round respectively, further regularizing the model and reducing the risk of variance inflation on the holdout test set.

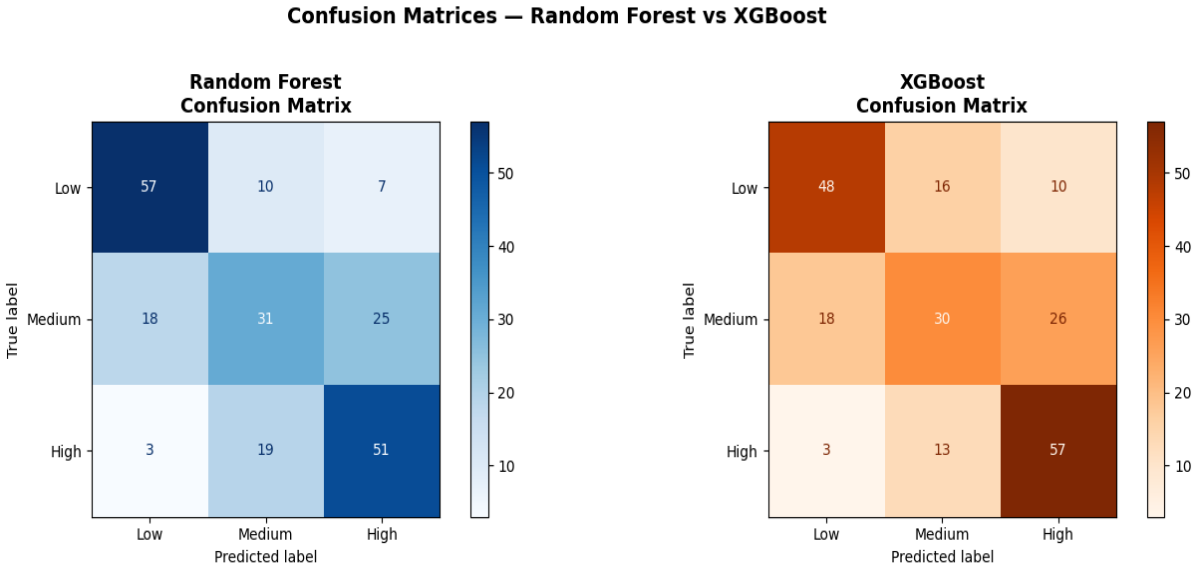
4.4 Model Performance and Evaluation Metrics

This section presents the performance metrics for the two ensemble machine learning models employed in this study, evaluating how Random Forest and XGBoost performed with respect to the classification of customers into Low, Medium, and High CLV tiers.

Table 2: Model Performance Comparison

Model	Accuracy	Precision	Recall	F1 Score
Random Forest	0.63	0.62	0.63	0.62
XGBoost	0.61	0.61	0.61	0.60

As presented in Table 2, Random Forest outperformed XGBoost across all evaluation metrics, achieving an accuracy of 63% and a weighted F1 score of .62 compared to XGBoost accuracy of 61% and F1 score of .60. Both models struggled most with the Medium CLV tier, reflecting boundary ambiguity between adjacent classes. Random Forest demonstrated more consistent tier-level classification and was accordingly selected as the final model for the subsequent CLV-to-product investment mapping analysis.



As depicted in Fig. 5, the confusion matrices corroborate the classification report findings. Random Forest correctly classified 57 Low, 31 Medium, and 51 High CLV customers, with the most notable misclassifications occurring within the medium tier, where 18 and 25 instances were incorrectly assigned to Low and High respectively. XGBoost correctly classified a higher number of High CLV customers (57) but performed comparatively weaker on the Low tier, correctly identifying only 48 instances against Random Forest's 57. Both models exhibited the greatest classification uncertainty within the medium tier, confirming that intermediate-value customers present the most challenging differentiation boundary for both ensemble approaches.

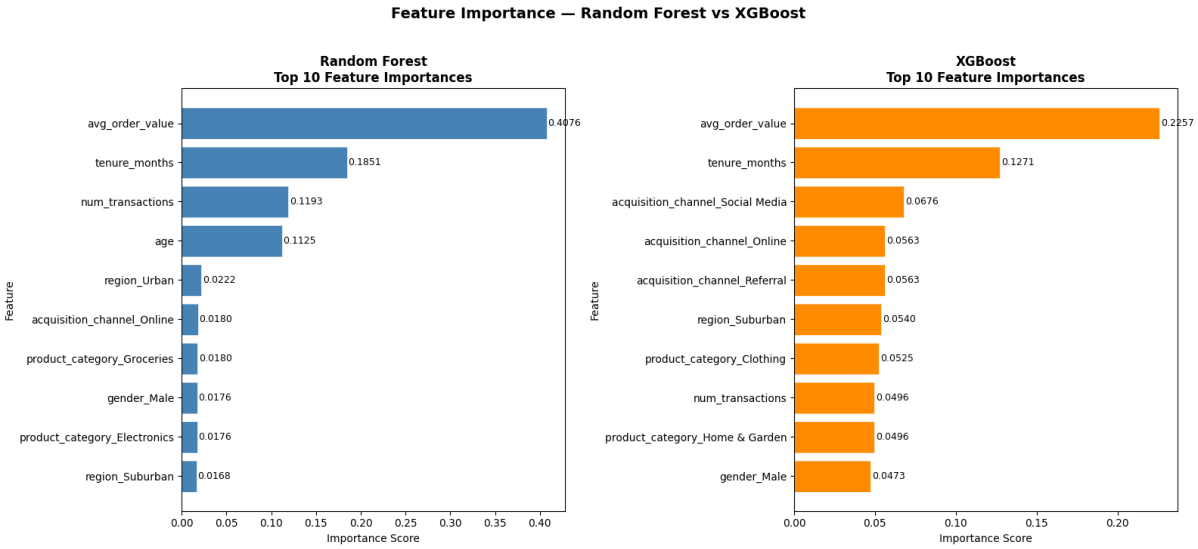


Figure 6: feature importance

As illustrated in Figure 6, both models consistently identified avg_order_value as the most influential predictor of CLV tier classification, recording importance scores of .4076 and .2257 for Random Forest and XGBoost respectively. Tenure_months ranked second in both models (.1851 and .1271), confirming that spending intensity and relationship longevity are the primary drivers of customer lifetime value. Random Forest additionally assigned notable importance to num_transactions (.1193) and age (.1125), while XGBoost distributed remaining importance more broadly across acquisition channels, region, and product categories, suggesting that contextual and behavioral variables contribute meaningfully to CLV classification beyond the dominant spending and tenure dimensions.

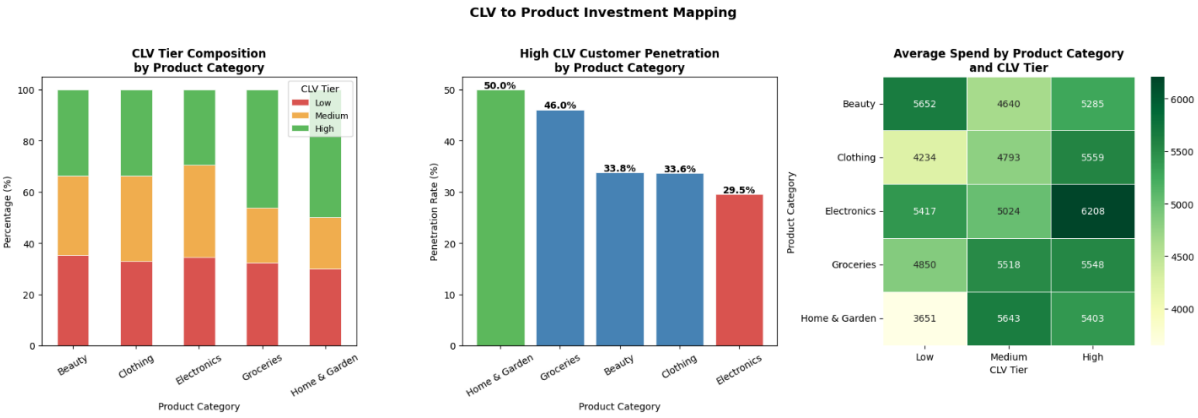


Figure 7: customer lifetime value to product investment mapping

The analysis of the CLV to product investment mapping as shown in Figure 7 indicated significant difference in the high value customer concentration of products. Home and Garden had the highest High CLV customer penetration rate (50.0%), and Groceries (46.0%), Beauty (33.8%), Clothing (33.6%), and Electronics (29.5%) are the categories with the highest concentration of the high-value customers. The average spend heatmap also indicated that the size of the High CLV customers produced the highest spend in Electronics (6,208), then in Groceries (5,548) and Clothing (5,559), indicating that, though High CLV customers are not occupied in the high numbers that Home and Garden and Groceries (penetration bright) categories, the customers that it does retain tend to spend the most individually and, as such, should be given significant product development consideration in addition to the previously dominant penetration categories.

4.6 Discussion of Findings

The key goal of this paper was to make Customer Lifetime Value predictions and look at how the predictions would be used to make product investment choices in a retail environment. The results all indicate that ensemble machine learning models, especially the Random Forest, can be successfully used to categorize customers into sensible CLV tiers, and the distributions of the resulting tiers have implications to product investment prioritization.

The identification of the two predictors as average order value and tenure months being the most dominant variables in predicting CLV tier in both models are in line with the recent empirical findings within the customer value literature. Chamberlain et al. (2021) also determined that the relationships between spending intensity and duration of relationship were the most robust predictors of long-term customer value in a variety of retail markets, and suggest that these two dimensions are more effective than other variables in capturing the compounding economic impact of customer relationships based on loyalty and high spending. These features bring both ensemble models into the same ball and further prove the validity of these models as predictors due to the convergence between the two methodologically different algorithms, lowering the chances that rankings of feature importance are an artefact of a single modelling algorithm.

The moderately good performance of both models in classification combined with the similarity in the inability to classify the Medium CLV tier perfectly is in line with the recent multi-class CLV classification experiments. Vanderveld et al. (2022) revealed similar trends of boundary ambiguity in between value segments, which the researchers attribute to the ambiguity between medium and adjacent tiers by the nature of the lack of longitudinal behavioral patterns. This implies that the predictive ceiling found in the current study is not specific to the models being used but instead a structural feature of CLV tier classification of cross-sectional retail data, which can be overcome more effectively by future studies that account using time-series purchasing behavior.

The CLV-to-product investment mapping results present the contribution with the greatest directly actionable results of the study. The high Home and Garden and Groceries category of High CLV customer penetration, therefore, support the stance taken by Tariq et al. (2023), who established that product category with an uneven distribution of high-value customers should be treated as a strategic retention resource, and is deserving of a prioritized development investment as compared to categories with low values-customer concentration. The opposite trend in the Electronics segment where the penetration to high-value, customers was smallest

yet the average spend per product category was the highest, is the opposite of the tendency towards high-value customer penetration and the consequent higher investment in these segmentations is the main focus of Benoit and Van den Poel (2021) who stated that with low-penetration, high-spend segments, the high value of customer penetration is unlikely to disperse across the board since the very core of these segmentations is their high level of spending intensity.

5. Conclusion and Recommendations

This paper evaluated Customer Lifetime Value by applying the ensemble machine learning models to make decisions about product investment in a retail setting. Based on a Kaggle-sourced dataset of 736 customers, the CLV tier prediction classifier was trained using the random forest and XGBoost classifiers, with the former showing a better overall performance. The analysis found that the predictors of customer lifetime value were average order value and tenure months, and the mapping of the customer lifetime value to product investment showed that the most concentration of high-value customers is in Home and Garden and Groceries categories. The overall findings in this report illustrate that structured CLV assessment offers organizations with a solid, information-based basis upon which to base the focus of product investment against the categories that are most likely to produce sustainable long term revenue growth.

Based on the learning gained in the research the following are recommended.

i. The retail organizations are advised to focus on investment in product development in Home and Garden and Groceries categories as they have a disproportionately large share of High CLV customers which generate revenue in the long-term.
ii. The companies must have measures that are specifically aimed at raising average order value and customer tenure since the two are the most strongly empirically supported predictors of high lifetime value classification.
iii. Future CLV modelling initiatives must include longitudinal and transactional data and purchase trajectory characteristics to overcome the current misclassification issue that is witnessed in the Medium CLV tier.
iv. The organizations are advised to establish specific retention measures on Beauty and Electronics segments whereby churn risk is high and high value customer penetration is relatively low despite high potential of individual spending.

References

1) Al Rafi, M., & Yassar, I. K. (2025). Forecasting Customer Lifetime Value: A Data-Driven Approach to Optimizing Marketing Budget Allocation. Journal of Computer Science and Technology Studies, 7(10), 537-550. doi:https://doi.org/10.32996/jcsts.2025.7.10.53
2) Ali, N., & Shabn, O. S. (2024). Customer lifetime value (CLV) insights for strategic marketing success and its impact on organizational financial performance. Cogent Business & Management, 11(1). doi:https://doi.org/10.1080/23311975.2024.2361321
3) Awaad, S. A., Kortam, W., & Ayad, N. (2024). Examining the impact of price sensitivity on customer lifetime value: empirical analysis. Cogent Business & Management, 11(1). doi:https://doi.org/10.1080/23311975.2024.2366441

4) Benoit, D. F., & Van den Poel, D. (2021). Benefits of quantile regression for the analysis of customer lifetime value in a contractual setting: An application in financial services. Expert Systems with Applications, 38(3), 10475–10483. https://doi.org/10.1016/j.eswa.2011.02.114
5) Blattberg, R. C., & Deighton, J. (1996). Manage marketing by the customer equity test. Harvard Business Review, 74(4), 136–144.
6) Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5–32. https://doi.org/10.1023/A:1010933404324
7) Chamberlain, B., Cardoso, Â., Liu, C., Pagliari, R., & Deisenroth, M. (2021). Customer lifetime value prediction using embeddings. Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, 186–196. https://doi.org/10.1145/3447548.3467120
8) Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 785–794. https://doi.org/10.1145/2939672.2939785
9) Cooper, R. G. (1990). Stage-gate systems: A new tool for managing new products. Business Horizons, 33(3), 44–54. https://doi.org/10.1016/0007-6813(90)90052-T
10) Coussement, K., Lessmann, S., & Verstraeten, G. (2017). A comparative analysis of data preparation algorithms for customer churn prediction in the telecommunications industry. Decision Support Systems, 95, 27–36. https://doi.org/10.1016/j.dss.2016.11.007
11) Cowan, G., Mercury, S., & Khraishi, R. (2023). Modelling customer lifetime-value in the retail banking industry. 1-23.
12) Fader, P. S., Hardie, B. G. S., & Lee, K. L. (2005). "Counting your customers" the easy way: An alternative to the Pareto/NBD model. Marketing Science, 24(2), 275–284. https://doi.org/10.1287/mksc.1040.0098
13) Gupta, S., & Lehmann, D. R. (2005). Managing customers as investments: The strategic value of customers in the long run. Wharton School Publishing.
14) Kumar, V., & Shah, D. (2009). Expanding the role of marketing: From customer equity to market capitalization. Journal of Marketing, 73(6), 119–136. https://doi.org/10.1509/jmkg.73.6.119
15) Ekinci, Y., Uray, N., & Ulengin, F. (2014). A customer lifetime value model for the banking industry: a guide to marketing actions. European Journal of Marketing, 48(4), 761-784. doi:https://doi.org/10.1108/EJM-12-2011-0714
16) Nyakeri, W. (2025). Strategic Customer Lifetime Value Prediction: Leveraging Machine Learning to Maximize Profitability in Retail -A Case Study Using 2010-2011 Online Retail Data. 1-13.
17) Oghenemaro, S. A. (2025). Optimizing Customer Lifetime Value (CLV) Prediction Models in Retail Banking Using Deep Learning and Behavioral Segmentation. American Journal of Humanities and Social Sciences Research, 9(7), 123-131.
18) Reinartz, W., & Kumar, V. (2003). The impact of customer relationship characteristics on profitable lifetime duration. Journal of Marketing, 67(1), 77–99. https://doi.org/10.1509/jmkg.67.1.77.18589
19) Rust, R. T., Lemon, K. N., & Zeithaml, V. A. (2004). Return on marketing: Using customer equity to focus marketing strategy. Journal of Marketing, 68(1), 109–127. https://doi.org/10.1509/jmkg.68.1.109.24030

20) Schmittlein, D. C., Morrison, D. G., & Colombo, R. (1987). Counting your customers: Who are they and what will they do next? Management Science, 33(1), 1–24. https://doi.org/10.1287/mnsc.33.1.1
21) Segarra-Moliner, J. R., & Bel-Oms, I. (2023). How Does Each ESG Dimension Predict Customer Lifetime Value by Segments? Evidence from U.S. Industrial and Technological Industries. Sustainability, 15(8), 6907. doi:https://doi.org/10.3390/su15086907
22) Tariq, M., Abbas, T., Abrar, M., & Iqbal, A. (2023). Does green product development and customer satisfaction influence green purchase intention? Journal of Retailing and Consumer Services, 71, 103–119. https://doi.org/10.1016/j.jretconser.2022.103199
23) Vanderveld, A., Pandey, A., Han, A., & Parekh, R. (2022). An engagement-based customer lifetime value system for e-commerce. Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2customer–2lifecycle. https://doi.org/10.1145/2939672.2939715
24) Wong, A., Garcia, A. V., & Lim, Y. W. (2025). A data-driven approach to customer lifetime value prediction using probability and machine learning models. Decision Analytics Journal, 16(1), 100601. doi:https://doi.org/10.1016/j.dajour.2025.100601
25) You, J. (2025). Customer Lifetime Value Forecasting Using Ensemble Learning on Ecommerce Big Data. International Conference on Digital Economy and Intelligent Computing, 49-53. doi:https://dl.acm.org/doi/10.1145/3746972.3746981