

# Customer Lifetime Value Prediction

DEVANSH MISHRA<sup>1</sup>, DEEPAM SINGH<sup>2</sup>, DR. ISHRAT ALI<sup>3</sup>, PROF. SANJAY PACHAURI<sup>4</sup>  
<sup>1, 2, 3, 4</sup>*Data Science, Greater Noida Institute of Technology*

## I. INTRODUCTION

In the contemporary business landscape, organizations are shifting from transactional metrics to a more relationship-oriented view of their customers. Central to this evolution is the concept of Customer Lifetime Value (CLV), a forward-looking prediction of the total net profit a business can expect from a customer over the entire duration of their relationship.<sup>4</sup> Unlike metrics that measure past performance, predictive CLV uses data to forecast future value, empowering businesses to make proactive, data-informed decisions regarding marketing spend, customer acquisition, and retention strategies.<sup>6</sup> The ability to accurately predict CLV allows a company to identify its most valuable customers, tailor its engagement strategies, and ultimately foster long-term, sustainable growth.<sup>9</sup>

This project outlines the development of a robust system to predict customer lifetime value by segmenting customers into distinct value-based archetypes. The core of this project is the application of machine learning techniques to transform raw transactional data into actionable business intelligence. By understanding which customers are likely to be the most valuable, businesses can optimize resource allocation, personalize marketing campaigns, and implement targeted retention efforts to reduce customer churn—the phenomenon of customers ceasing their relationship with a company.<sup>10</sup> Given that acquiring a new customer can be significantly more expensive than retaining an existing one, predictive CLV modeling is a critical capability for maintaining profitability and a competitive edge.<sup>12</sup>

specifically focusing on predictive analytics within the domain of marketing and customer relationship management (CRM).<sup>14</sup> The project leverages a synergistic combination of unsupervised and supervised machine learning models to achieve its goal. The primary technologies and technical concepts involved include:

- **RFM (Recency, Frequency, Monetary) Analysis:** A foundational technique in marketing analytics used for feature engineering. It quantifies customer purchasing behavior by analyzing how recently a customer has purchased (Recency), how often they purchase (Frequency), and how much they spend (Monetary value).<sup>16</sup> This method transforms complex transaction histories into a simple, standardized format.

RecencyCluster	count	mean	std	min	25%	50%	75%	max
0	9540	77.679245	22.850898	48.0	59.0	72.5	93.0	131.0
1	4780	304.393305	41.183489	245.0	266.25	300.0	336.00	373.0
2	5680	184.625000	31.753602	132.0	156.75	184.0	211.25	244.0
3	1950	17.488205	13.237058	0.0	6.00	16.0	28.00	47.0

- **K-Means Clustering:** An unsupervised machine learning algorithm used to partition the customer base into distinct groups or "clusters" based on their RFM scores. This process identifies naturally occurring segments of customers with similar behaviors (e.g., high-spenders, new customers, at-risk customers) without any predefined labels.<sup>7</sup>

The field of this project is applied data science,

count	mean	std	min	25%	50%	75%	max	FrequencyCluster
0	3496.0	49.525744	44.954212	1.0	15.0	33.0	73.0	190.0
1	429.0	331.221445	133.856510	191.0	228.0	287.0	399.0	803.0
2	22.0	1313.136364	505.934524	872.0	988.5	1140.0	1452.0	2782.0
3	3.0	5917.666667	1805.062418	4642.0	4885.0	5128.0	6555.5	7983.0

- XGBoost (Extreme Gradient Boosting): A powerful and efficient supervised machine learning algorithm used for classification and regression tasks. In this project, XGBoost is used to build a multi-class classification model. The model is trained on the customer segments identified by the K-Means algorithm to predict the value segment to which any new or existing customer belongs.<sup>4</sup> XGBoost is known for its high performance, speed, and scalability.<sup>18</sup>

By integrating these techniques, this project will deliver a comprehensive, end-to-end framework for CLV prediction. The outcome is not just a static analysis but a dynamic predictive engine that can be operationalized to drive smarter, more profitable customer engagement strategies.

## II. OBJECTIVES

Based on the strategic importance of CLV and the capabilities of modern machine learning, this project aims to achieve the following objectives:

- To develop a robust data processing pipeline to calculate Recency, Frequency, and Monetary (RFM) scores from raw customer transaction data, effectively quantifying purchasing behavior.
- To apply unsupervised machine learning, specifically the K-Means clustering algorithm, to segment the customer base into distinct, value-based archetypes based on their RFM scores.
- To build and train a supervised multi-class classification model using the XGBoost algorithm to predict the CLV segment for any

given customer.

- To rigorously evaluate the performance of the predictive model using standard classification metrics, such as precision, recall, and F1-score, to ensure its accuracy and reliability for business applications.
- To establish a methodological framework that can be operationalized by a business to automate customer segmentation and inform targeted marketing strategies for different customer groups, thereby enhancing customer retention and maximizing profitability.

## III. METHODOLOGY/ PLANNING OF WORK

The project will be executed following a structured, multi-stage methodology that combines data engineering, unsupervised learning, and supervised predictive modeling.

### 1. Data Preparation and Feature Engineering (RFM Analysis)

The initial step involves transforming raw transactional data into meaningful features that capture customer behavior. This will be achieved using RFM analysis. For each customer, three key metrics will be calculated:

- Recency (R): The time elapsed since the customer's last purchase.
- Frequency (F): The total number of purchases made by the customer.
- Monetary (M): The total monetary value of all purchases made by the customer.

These raw values will then be normalized using a quintile-based scoring system, assigning each customer a score from 1 to 5 for each of the three dimensions. This process condenses each customer's entire purchase history into a simple and powerful three-dimensional feature vector.

	Recency	Frequency	Revenue
Overall Score			
0	304.584388	21.995781	303.339705
1	185.362989	32.596085	498.087546
2	78.972856	47.060803	871.842586
3	20.662252	68.374172	1089.271213
4	14.913043	271.678930	3609.566689
5	9.585034	374.136054	9169.540884
6	7.740741	876.037037	22777.914815
7	1.857143	1272.714286	103954.025714
8	1.333333	5917.666667	42177.930000

## 2. Unsupervised Customer Segmentation (K-Means Clustering)

With the RFM scores calculated, the next step is to identify natural groupings within the customer base. The K-Means clustering algorithm will be applied to the three-dimensional RFM data. A key parameter for K-Means is the number of clusters (K). The optimal value for K will be determined using the Elbow Method, which involves plotting the Within-Cluster Sum of Squares (WCSS) for a range of K values and identifying the "elbow" point where the rate of decrease slows significantly.<sup>24</sup> Once the optimal K is determined, the algorithm will be run to assign each customer to one of the K clusters. These clusters will then be profiled by analyzing their average RFM scores to create descriptive personas (e.g., "Champions," "At-Risk," "New Customers").

## 3. Supervised Predictive Modeling (XGBoost)

The cluster labels generated by the K-Means algorithm will serve as the target variable for a supervised learning model. The goal is to train a model that can predict the CLV segment of a customer based on their RFM scores. For this task, the XGBoost (Extreme Gradient Boosting) algorithm will be used due to its high performance and scalability.<sup>30</sup> The dataset, consisting of customer RFM scores (features) and their assigned cluster labels (target), will be split into a training set and a testing set. The XGBoost classifier will be trained on the training data to learn the patterns that link RFM characteristics to specific value segments.

## 4. Model Evaluation

The performance of the trained XGBoost model will be evaluated on the unseen testing set. A classification report will be generated to assess its predictive accuracy across each customer segment. The key metrics for evaluation will be:

- Precision: The proportion of correct positive predictions.
- Recall: The proportion of actual positives that were correctly identified.
- F1-Score: The harmonic mean of precision and recall, providing a balanced measure of performance.<sup>34</sup>

This rigorous evaluation will determine the model's reliability and its readiness for deployment in a business context.

## REFERENCES

- [1] A. Hughes, *Strategic Database Marketing*. Chicago, IL, USA: Probus Publishing Co., 1994.
- [2] J. MacQueen, "Some methods for classification and analysis of multivariate observations," in *Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability*, 1967, vol. 1, pp. 281–297.<sup>35</sup>
- [3] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Francisco, CA, USA, Aug. 2016, pp. 785–794.<sup>18</sup>

## WORKS CITED

- [1] Analysis and Optimization of Customer Lifetime Value Prediction using Machine Learning and Deep Learning Models by RFM Techniques, accessed September 17, 2025, [https://ijwr.usc.ac.ir/article\\_221737\\_7c7d968f0ba26030f74c4f707cabd7ad.pdf](https://ijwr.usc.ac.ir/article_221737_7c7d968f0ba26030f74c4f707cabd7ad.pdf)
- [2] Ultimate Guide to CLV Prediction with ML - M ACCELERATOR by M ..., accessed September 17, 2025, <https://maccelerator.la/en/blog/entrepreneurship/ultimate-guide-to-clv-prediction-with-ml/>
- [3] Why You Need to Predict Customer Lifetime Value | Pecan AI, accessed September 17, 2025, <https://www.pecan.ai/blog/why-predict-customer-lifetime-value/>
- [4] Customer Churn Prediction using Machine Learning Approach: A Comprehensive Study, accessed September 17, 2025, [https://www.researchgate.net/publication/390479920\\_Customer\\_Churn\\_Prediction\\_using\\_Machine\\_Learning\\_Approach\\_A\\_Comprehensive\\_Study](https://www.researchgate.net/publication/390479920_Customer_Churn_Prediction_using_Machine_Learning_Approach_A_Comprehensive_Study)
- [5] A Survey on Customer Churn Prediction using Machine Learning Techniques, accessed September 17, 2025, <https://www.ijcaonline.org/archives/volume154/number10/kumar-2016-ijca-912237.pdf>
- [6] CUSTOMER CHURN PREDICTION - Kaggle, accessed September 17, 2025, <https://www.kaggle.com/code/bhartiprasad17/customer-churn-prediction>
- [7] Bank Customer Churn Prediction Using Machine Learning - Analytics Vidhya, accessed September 17, 2025,

<https://www.analyticsvidhya.com/blog/2022/09/bank-customer-churn-prediction-using-machine-learning/>

- [8] Applications of Machine Learning (ML) in the context of marketing: a bibliometric approach, accessed September 17, 2025, <https://pmc.ncbi.nlm.nih.gov/articles/PMC11926528/>
- [9] Machine Learning In Marketing: Advantages, Applications & Examples, accessed September 17, 2025, <https://edvancer.in/machine-learning-in-marketing-applications/>
- [10] (PDF) Research on customer lifetime value based on machine learning algorithms and customer relationship management analysis model - ResearchGate, accessed September 17, 2025, [https://www.researchgate.net/publication/368484375\\_Research\\_on\\_customer\\_lifetime\\_value\\_based\\_on\\_machine\\_learning\\_algorithms\\_and\\_customer\\_relationship\\_management\\_analysis\\_model](https://www.researchgate.net/publication/368484375_Research_on_customer_lifetime_value_based_on_machine_learning_algorithms_and_customer_relationship_management_analysis_model)
- [11] 3 Best Machine Learning Models to Predict Customer Lifetime Value (CLTV), accessed September 17, 2025, <https://blueorange.digital/blog/3-best-machine-learning-models-to-predict-customer-lifetime-value-cltv/>
- [12] XGBoost: A Scalable Tree Boosting System, accessed September 17, 2025, <https://arxiv.org/abs/1603.02754>
- [13] XGBoost: A Scalable Tree Boosting System, accessed September 17, 2025, <https://www.kdd.org/kdd2016/papers/files/rfp0697-chenAemb.pdf>
- [14] XGBoost: A Scalable Tree Boosting System - ResearchGate, accessed September 17, 2025, [https://www.researchgate.net/publication/310824798\\_XGBoost\\_A\\_Scalable\\_Tree\\_Boosting\\_System](https://www.researchgate.net/publication/310824798_XGBoost_A_Scalable_Tree_Boosting_System)