Predicting Customer Segment Changes to Enhance Customer Retention: A Case Study for Online Retail using Machine Learning

Lahcen ABIDAR, Dounia ZAIDOUNI, Ikram EL ASRI and Abdeslam ENNOUAARY
Department of Mathematics, Networks and Computer Science
National Institute of Posts and Telecommunications, Rabat, Morocco

***Abstract*—In today's highly competitive marketplace, advertisers strive to tailor their messages to specific individuals or groups, often overlooking their most significant clients. The Pareto principle, asserting that 80% of sales come from 20% of customers, offers valuable insights, imagine if companies could accurately forecast this vital 20% and recognize its historical significance. Predicting customer lifetime value (CLV) at this juncture becomes crucial in aiding firms to effectively prioritize their efforts. To achieve this, organizations can leverage predictive models and analytical tools to target specific customers with tailored campaigns, enabling well-informed decisions about advertising investments. By being aware of these segment transitions, advertisers can efficiently deploy resources and increase their return on investment. By implementing the strategies outlined in this study, businesses can gain a competitive edge by identifying and retaining their most valuable clients. The potential for growth and client retention is immense when anticipating changes in customer segments and adjusting advertising strategies accordingly. This paper provides a comprehensive methodology, tools, and insights to assist marketers in optimizing their advertising campaigns by anticipating customer lifetime value and actively predicting changes in client segmentation.**

***Keywords*—Customer segment changes; customer retention; marketing actions; informed decisions; advertising strategies**

I. Introduction

In today's highly competitive marketplace, advertisers face the ongoing challenge of delivering targeted and personalized advertisements to capture the attention of potential customers. However, amidst the quest for broad reach and mass appeal, businesses often overlook a crucial aspect – focusing on their most valuable customers. The Pareto principle [1], a well-known economic principle, sheds light on this phenomenon by revealing that a small portion of customers typically contributes a significant proportion of sales. While the Pareto principle has long been cited in reference to sales patterns, its implications for advertising strategies have been under-explored. What if businesses could not only identify this 20% historically but also predict it for the future? The concept of predicting customer lifetime value (CLV) arises as a powerful tool in this context. By leveraging predictive modeling and analytical techniques, businesses can forecast future customer behavior and identify those individuals who will likely make up the high-value customer segment. The primary objective of this paper is to propose an approach for predicting customer segment changes based on CLV predictions. By accurately predicting shifts in customer segments, businesses can strategically prioritize their marketing actions and allocate resources more efficiently. This includes determining the optimal investment in advertising, identifying the specific customers to target with tailored campaigns, and devising strategies to transition customers from one segment to another. Understanding and leveraging customer segment changes present several strategic advantages for businesses. Firstly, it allows for the efficient allocation of advertising resources [2], ensuring that marketing efforts are focused on the most valuable customers who are likely to drive significant sales [3]. Secondly, by tailoring advertising messages and offers to this high-value segment, businesses can improve customer engagement and conversion rates[4]. Lastly, actively managing customer transitions between segments enables businesses to nurture relationships, increase customer loyalty [5], and maximize long-term customer value. This paper serves as a comprehensive guide for businesses seeking to optimize their advertising strategies by harnessing the power of CLV predictions and proactively targeting customer segments. Through an exploration of predictive modeling techniques and actionable insights derived from customer segment changes, businesses can gain a competitive edge in today's dynamic marketplace. By aligning their advertising efforts with anticipated shifts in customer segments, businesses can enhance customer retention, maximize profitability, and foster sustainable growth. This paper is organized as follows: In Section II, we provide an overview of the existing literature. Next, in Section III, we present a comprehensive framework that outlines the step-by-step process of predicting customer segment changes based on customer lifetime value. Moving on to Section IV, we present the empirical study conducted to evaluate the effectiveness of our approach. We describe the dataset used, the experimental setup, and the evaluation metrics employed. We then present and analyze the results obtained from applying the proposed approach, discussing the performance of the predictive models and any significant findings or insights gained. In Section V, we interpret and discuss the empirical results in the context of our research objectives and identify potential areas for future research and improvement. Finally, in Section VI, we summarize the key findings of our study and draw meaningful conclusions based on the empirical analysis and discussions. By structuring the paper in this manner, we aim to provide a comprehensive overview of our research, methodology, and findings, offering valuable insights and practical guidance for advertisers seeking to optimize their advertising strategies using customer lifetime value prediction and proactive customer segment targeting.

II. Literature Review

In the realm of personalized advertising and maximizing return on investment, the understanding of customer segments and their dynamic changes over time has emerged as a crucial area of research. This literature review explores key studies related to customer segment changes, customer lifetime value (CLV) prediction, and proactive customer segment targeting.

Customer segmentation, a vital concept in marketing, allows businesses to categorize their customer base into distinct groups based on common characteristics [6], behaviors, or preferences. Various factors influencing customer segment changes and transitions have been examined by researchers. Christy et al highlighted the significance of RFM analysis for identifying valuable customer segments and guiding marketing initiatives [7]. Yuliari et al introduced a customer segmentation method using fuzzy C-means and fuzzy RFM, accounting for uncertainties in customer data [8]. Sembiring Brahmana et al investigated customer segmentation using the RFM model and clustering techniques such as K-means, K-medoids, and DBSCAN [9]. Dullaghan and Rozaki explored machine learning techniques for dynamic customer segmentation analysis in the mobile industry [10]. Ahani et al conducted market segmentation and travel choice prediction in spa hotels using online reviews [11]. Albuquerque et al applied support vector clustering for customer segmentation in the context of mobile TV service [12].

The prediction of customer lifetime value (CLV) has gained significant attention as it enables businesses to identify their most valuable customers historically and forecast their future value. Several predictive modeling and analytical techniques have been explored in this context. De Marco et al utilized cognitive analytics and artificial neural networks to manage CLV, facilitating customer value prediction and optimization. They found that the self-organizing map better classifies the customer base of the retailer [13]. Marisa et al explored the relationship between CLV and core drives, using clustering and the octalysis gamification framework. Their study analyzed the relationship between CLV and eight core drives of customer motivation [14]. Yuan et al focused on a data-driven customer segmentation strategy based on the contribution to system peak demand [15]. Mosaddegh et al studied the dynamics of bank customers through value segments using big data analytics, identifying six major categories, including the pattern of Local Leaders whose transitions are repeated by some follower groups within the next two periods [16]. Khalili-Damghani et al proposed a hybrid approach combining clustering, rule mining, and decision tree analysis for personalized marketing [17].

To optimize advertising strategies, businesses need to proactively target specific customer segments that are likely to yield higher returns. Researchers have developed approaches to identify and prioritize these segments. Heldt et al introduced a predictive model called RFM/P, extending RFM analysis to enhance customer segmentation and targeting strategies [18]. Abidar et al proposed a new strategy for customer segmentation using machine learning techniques, highlighting the importance of targeted actions in marketing. Their approach demonstrates the effectiveness of machine learning in identifying customer segments and enabling businesses to tailor their marketing efforts for improved customer satisfaction and profitability [19]. Yuan et al focused on a data-driven customer segmentation strategy based on the contribution to system peak demand [15].

The ability to predict customer segment changes and align advertising efforts accordingly present substantial opportunities for growth and customer retention. By effectively identifying and engaging their most valuable customers, businesses can gain a competitive edge in the marketplace. The findings from studies in this field highlight the importance of leveraging CLV prediction and proactive customer segment targeting to optimize advertising strategies, allocate resources efficiently, and maximize return on investment.

In summary, the literature review emphasizes the significance of understanding customer segment changes, predicting CLV, and leveraging proactive targeting strategies in the realm of personalized advertising. The studies reviewed provide valuable insights and methodologies for businesses seeking to optimize their advertising strategies and enhance customer retention.

III. Workflow Model

In order to reach a final outcome, various techniques and methods will be utilized in this research. The resulting framework workflow, depicted in Fig. 1, incorporates customer segmentation, RFM parameters, clustering, data analytics, CLV Prediction, and targeted actions.

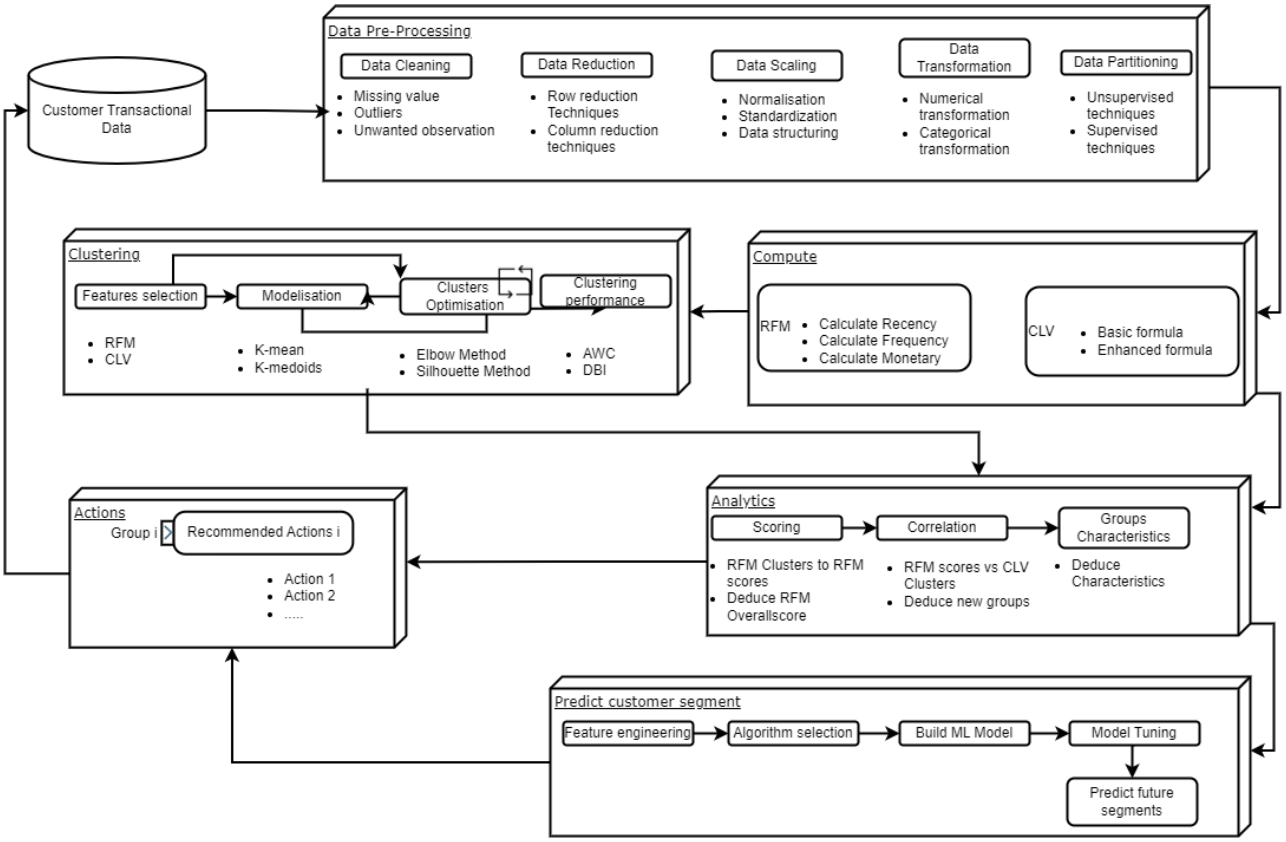
A. Data Pre-processing

1) Data cleaning: When creating operational data, there are two standard approaches to handling missing numbers. As most data mining algorithms cannot handle data with missing values, the initial step is to simply remove data samples with missing values. This approach is only appropriate when the percentage of missing values is negligible. The second is to use missing value imputation techniques to substitute inferred values for missing data. There are two techniques for detecting outliers: statistical and clustering-based techniques [20].

2) Data reduction: Row-wise for data sample reduction and column-wise for data variable reduction are the two usual directions in which data reduction is carried out. Row-wise data reduction is possible using a variety of data sampling methods, including random and stratified sampling. The goal of feature extraction is to create new features based on linear or nonlinear combinations of existing variables, as opposed to feature selection, which only chooses usable features from already-existing variables [20].

3) Data scaling: Predictive modeling frequently requires data scaling, particularly when the input variables have multiple scales. The max-min normalization (i.e., $x' = x - x_{min}/x_{max} - x_{min}$) and z-score standardization (i.e., $x' = x - \mu/\sigma$) are two of the most widely used methods in the building field, where x_{min} and x_{max} refer to the minimum and maximum of variable x, values of the variable, μ is the mean and σ is the standard deviation [20].

4) Data transformation: Data transformation is mostly used in the construction industry to convert numerical data into categorical data in order to assure interoperability with data mining methods. Due to their simplicity, the equal-width and equal-frequency approaches are frequently utilized [20].



(a)

Fig. 1. Framework workflow.

5) Data partitioning: The goal of data partitioning is to separate the entire set of data into various categories for in-depth study. For this goal, decision tree approaches and clustering analysis have been frequently applied in the building industry. For data partitioning, a variety of clustering methods have been used, including fuzzy c-means clustering, hierarchical clustering, entropy weighting k-means (EWKM), and k-means [20].

B. Clustering

1) Features selection: In practice, it is uncommon for all of a dataset's variables to be helpful in creating a machine learning model. Repetitive variables decrease a model's capacity to generalize and may also lower a classifier's overall accuracy. A model's overall complexity is also increased by including more variables. The objective of feature selection in machine learning is to identify the best set of features that make it possible to create effective models of the phenomena being examined. In machine learning, there are two different kinds of feature selection methods: supervised and unsupervised methods.

2) Modelization: A file that has been trained to detect various patterns is referred to as a machine learning model. By giving a model a method it can use to analyze and learn from a set of data, we may train it on that data. After the model has been trained, we can use it to analyze new data and forecast what will happen to it.

3) Cluster optimisation: Every clustering algorithm has its own strengths and weaknesses, In order to overcome these flaws in clustering algorithms, it is necessary to estimate the number of clusters based on assumptions and rely significantly on the initial centroids choice. It is vital to optimize, and the Elbow approach is one of the most used cluster optimization techniques.

4) Clustering performance: Any typical clustering system must answer the fundamental question of how accurate or reliable the clustering is. The separation between clusters is calculated using the Silhouette Score and Silhouette Plot. It shows the distance between each point in a cluster and points in other clusters. This metric, which has a range of [-1, 1], is excellent for visually examining similarities and differences between clusters.

C. Compute

1) RFM compute: RFM is a technique for providing significant value to each consumer. It is mostly utilized in marketing and has drawn the attention of the retail and business services industries. Based on the following criteria, RFM:

- Recent: When was the client's most recent order?

- Frequency: How frequently do they purchase?
- Monetary: How much do they spend?

The Fig. 2 displays a series of prior sales for a group of four clients.

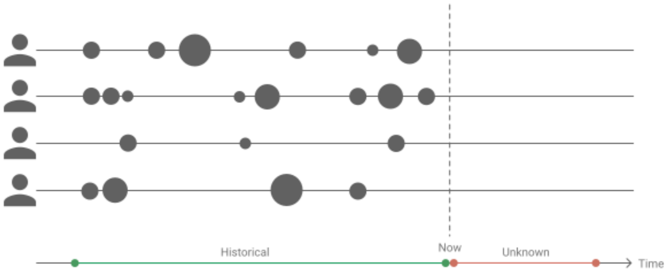


Fig. 2. Past sales for a set of four customers.

The diagram depicts the RFM values for the clients, with the following information for each client:

- Recency: The amount of time that has passed since the last purchase, as indicated by the separation between the rightmost circle and the vertical dotted line that reads ‘Now”.
- Frequency: The space between the circles on a single line, which represents the interval between purchases.
- Monetary: The size of the circle represents the amount of money spent on each purchase. This sum could represent either the typical order value or the number of products the buyer ordered.

2) CLV compute: CLV is calculated as the total of net cash flows from consumers over their anticipated lifetime, taking the time value of money into account. The following formula can be used to represent this model. The research in [21] and Table I is showing it’s parameters:

$$CLV = \sum_{i=1}^{n} \frac{R_i - C_i}{(1+d)^{i-0.5}} \tag{1}$$

TABLE I. CLV FORMULA PARAMETERS

Var	Explanation	Operationalization
n	Expected life of a customer	n = the total number of periods of projected life of the customer under consideration
C_i	The total cost of customer in period i	Total cost of generating the revenue R_i in period i
R_i	Total revenue of customer in period i	The revenues of customers were assigned as their monetary values.
d	Discount rate (annual)	Discount.

D. Analytics

The analysis phase includes several key components. One of these components is the RFM cluster analysis, which is used to assign RFM scores to different customer segments. The RFM scores represent the recency, frequency, and monetary value of customer transactions, providing insights into their purchasing behavior. the overall score is calculated based on the RFM scores, this overall score serves as a valuable metric for evaluating customer segments. To enhance understanding, the analysis also includes a detailed examination of the characteristics and attributes of each customer group within the RFM and CLV clusters.

E. Predict Customer Segment

In this part, we focus on the process of predicting future customer segments based on the developed workflow model. We delve into the various steps involved, including feature engineering, algorithm selection, building the machine learning model, model tuning, and ultimately predicting the future segments of customers. Feature engineering plays a crucial role in creating meaningful predictors for the machine learning model. We explore the different techniques and strategies employed to transform raw data into informative features that capture the relevant characteristics of customer behavior. Choosing the appropriate machine learning algorithm is a critical decision that impacts the accuracy and effectiveness of segment prediction. We explore a range of algorithms commonly used in customer segmentation tasks, such as decision trees, random forests, logistic regression, and gradient-boosting algorithms. In building the ML Model section, we detail the process of building the machine learning model for predicting customer segments. We discuss the steps involved in model training, validation, and evaluation. We split the dataset into training and test sets, and we discuss model performance metrics and interpretability, ensuring that the chosen model aligns with the objectives and requirements of segment prediction. In the Model Tuning section, we did some tuning to optimize the performance of the machine learning model. The final step in this part of our workflow is to use the trained and tuned model to predict future customer segments.

F. Actions

Based on the results obtained, businesses can implement effective actions to help to develop customer retention strategies and allocate their advertising and marketing investments more strategically. They can also tailor their advertising campaigns to target specific customer groups and help to draw roadmap for segment transition planning.

IV. EMPIRICAL RESULTS AND ANALYSIS

A. Data

The data used in this study was gathered from an online retailer [22]. The dataset covers the time period from the end of 2009 to November 2011. The collection includes 16759 invoices for 3881 items produced by 889 clients (Table II).

B. Data Pre-processing

1) Data cleaning: Following the completion of the data cleaning process, certain incorrect and missing values were removed from the data set. Table III summarises The attributes that were employed in this study. Table IV demonstrates a few modifications that we make to the data to make it cleaner.

TABLE II. TRANSACTIONAL DATA

Id	Invoice	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country
931932	574387	22726	ALARM CLOCK BAKELIKE GREEN	8	11/4/2011 11:04	3.75	12944.0	United Kingdom
404468	528600	22028	PENNY FARTHING BIRTHDAY CARD	12	10/22/2010 14:57	0.42	12787.0	Netherlands
764721	560569	22423	REGENCY CAKESTAND 3 TIER	1	7/19/2011 14:04	12.75	12480.0	Germany
442525	532056	21530	DAIRY MAID TOASTRACK	6	11/10/2010 14:27	2.95	12739.0	United Arab Emirates
272695	516189	85049D	BRIGHT BLUES RIBBONS	12	7/18/2010 15:56	1.25	12625.0	Germany
164614	505168	16053	POPART COL BALLPOINT PEN ASST	50	4/20/2010 12:48	0.21	14156.0	EIRE
88489	497879	21931	JUMBO STORAGE BAG SUKI	10	2/14/2010 11:15	1.95	12422.0	Australia

TABLE III. ATTRIBUTES

Attributes	Description
InvoiceNo	Unique ID to identify each Invoice
StockCode	Unique ID for each item in stock
Description	A short description for each item
Quantity	Number of items bought
UnitPrice	The price of each item
CustomerID	Unique ID for each Customer
Country	The country where the Customer lives

TABLE IV. DATA CLEANING

Problem	Solution
Null Invoices	Subtract from the dataset (not pertinent to this study)
Negative UnitPrice	Delete from this data (the organization included this in order to adjust bad credit)
Invoice with no customerID	Since we will be doing customer segmentation, remove any rows where customerID is NA.

We use whole prepared population in the analysis. Thus, we did not use any sampling method.

2) Data selection: We chose the following elements for this study: CustomerID, InvoiceDate, Quantity, UnitPrice. These attributes will make it easier for us to apply RFM models to this company's customers and determine customer lifetime value.

3) Data transformation: No data transformations have been made in the database.

C. Time Frame for CLTV Calculation

In this study, we will assess the Customer Lifetime Value (CLTV) over a 6-month period and use it in parameter correlation analysis with other variables for the purpose of feature engineering.

D. LTV Clusters

The Table V represents the three LTVCluster derived from the Elbow method, along with various statistical measures such as count, mean, standard deviation, minimum, 25^{th} percentile, median (50^{th} percentile), 75^{th} percentile, and maximum. Here's what each column represents: **LTVCluster** represents different clusters or segments based on the Lifetime Value (LTV) of customers and the **Count** column indicates the number of data points or observations within each LTVCluster. The **Mean** column represents the average value of the Lifetime Value within each LTVCluster. It provides insight into the average LTV for customers within each cluster. The **Std** column represents the standard deviation of the Lifetime Value within each LTVCluster. It provides a measure of the variability or dispersion of LTV values within each cluster and the **Min** column indicates the minimum value of the Lifetime Value within each LTVCluster. It represents the lowest observed LTV for customers within each cluster. The 25^{th} percentile column represents the value below which 25% of the Lifetime Values fall within each LTVCluster. It provides an insight into the lower quartile or first quartile value for LTV within each cluster and The 50^{th} percentile column represents the median value of the Lifetime Value within each LTVCluster. It indicates the midpoint of the LTV distribution within each cluster. The 75^{th} percentile column represents the value below which 75% of the Lifetime Values fall within each LTVCluster. It provides an insight into the upper quartile or third quartile value for LTV within each cluster. Finally, the **Max** column indicates the maximum value of the Lifetime Value within each LTVCluster. It represents the highest observed LTV for customers within each cluster.

These statistics provide an overview of the distribution and characteristics of the Lifetime Value within each LTVCluster. They can be used to compare and understand the differences in LTV between different customer clusters or segments.

E. Feature Engineering

In the feature engineering phase, we utilize RFM (Recency, Frequency, Monetary) scores calculated using the model introduced in our previous paper. These scores are merged with the calculated Customer Lifetime Value (CLV) for the 6-month period. To prepare the data for modeling, we perform various feature engineering techniques. First, we convert categorical columns, such as segment categories (low, mid, high), into numerical columns by assigning them values of 0 or 1. This enables us to incorporate these categorical variables into our machine-learning model effectively. Next, we examine the correlation between the features and our target variable, LTVCluster (Table VI). The correlation coefficients are as follows:

- LTVCluster: 1.000000
- DataF2_Monetary: 0.861441
- Monetary: 0.578009
- MonetaryCluster: 0.505388
- Segment_High-Value: 0.450551
- Frequency: 0.406573
- FrequencyCluster: 0.406352
- OverallScore: 0.392761
- RecencyCluster: 0.231834

TABLE V. LTV CLUSTER

LTVCluster	Count	Mean	Std	Min	25%	50%	75%	Max
0	2014.0	332.786480	390.273105	-609.40	0.0000	191.805	569.8225	1369.27
1	400.0	2408.376375	917.844417	1375.75	1695.0325	2075.775	2905.2150	4969.83
2	50.0	7633.641200	2225.110687	5074.93	6088.2675	6708.085	8838.1825	13636.42

- Segment_Mid-Value: 0.105407
- CustomerID: -0.055825
- Recency: -0.241712
- Segment_Low-Value: -0.263938

These correlation values provide insights into the relationships between the LTVCluster and other parameters. A positive correlation indicates a direct relationship, where an increase in one parameter is associated with an increase in LTVCluster. For example, DataF2_Monetary, Monetary, and MonetaryCluster show strong positive correlations, suggesting that higher monetary value and overall customer spending are indicative of a higher LTVCluster. Conversely, negative correlation coefficients suggest an inverse relationship, where an increase in one parameter is associated with a decrease in LTVCluster. In this case, Recency and Segment_Low-Value exhibit negative correlations, indicating that longer periods of inactivity and lower segment values are linked to a lower LTVCluster.

To build the machine-learning model, we split the dataset into training and test sets. The training set is utilized for training the model, while the test set is used to evaluate the model's performance on unseen data.

F. Algorithms Comparison

In this section, we compare the performance of different machine learning algorithms based on their mean and standard deviation scores in Table VII.

Among the algorithms evaluated in our study, LogisticRegressionCV (LR) demonstrated a mean score of 0.839304 and a low standard deviation of 0.019146, indicating consistently good performance. The XGBClassifier (XGB) followed closely with a mean score of 0.831148 and a standard deviation of 0.012742, showcasing reliable and consistent results. The KNeighborsClassifier (KNN) achieved a mean score of 0.838491, similar to LR, but with a slightly higher standard deviation of 0.023427, implying a slightly higher variability in its performance. On the other hand, the DecisionTreeClassifier (DT) algorithm outperformed the others with the highest mean score of 0.845802 and a low standard deviation of 0.012548, demonstrating both high accuracy and consistency. The RandomForestClassifier (RF) achieved a mean score of 0.831974, similar to XGB and LR, with a moderate standard deviation of 0.015155. The AdaBoostClassifier (ADA) also performed well with a mean score of 0.834416 and a low standard deviation of 0.014393, comparable to XGB and LR. Lastly, the SVC algorithm obtained a mean score of 0.825506, slightly lower than other algorithms, but with a low standard deviation of 0.012568, indicating consistent results. These findings provide valuable insights into the performance and stability of each algorithm, guiding the selection of the most suitable model for predicting customer segment changes.

Based on this analysis, the DecisionTreeClassifier (DT) and XGBClassifier (XGB) show the highest mean score and lowest standard deviation, suggesting it performs the best among the listed algorithms. However, it's also important to consider other factors such as computational complexity, interpretability, and specific requirements of your task when choosing the most suitable algorithm.

The choice between XGBClassifier and DecisionTreeClassifier depends on various factors and considerations. Our preference for XGBClassifier stems from its utilization of the powerful XGBoost algorithm, renowned for its exceptional performance in machine learning tasks. Unlike a single Decision Tree, XGBClassifier excels in handling complex datasets and often achieves higher accuracy. This is achieved by combining multiple weak decision trees through boosting techniques, resulting in improved overall performance. While Decision trees can be prone to overfitting, XGBClassifier incorporates regularization techniques such as shrinkage to mitigate this issue. Moreover, it performs automatic feature selection, ensuring the inclusion of relevant features and reducing the risk of using irrelevant or noisy ones. Although decision trees are generally regarded as more interpretable, XGBClassifier provides valuable insights through variable importances, indicating the relative significance of features. Additionally, XGBClassifier offers greater flexibility in handling missing values, enhancing the robustness of the model. With its ability to capture complex nonlinear relationships and interactions through gradient boosting, XGBClassifier surpasses DecisionTreeClassifier in scenarios where features exhibit nonlinear relationships with the target variable. Furthermore, XGBClassifier's optimization for performance and efficiency, including parallel processing and tree pruning techniques, makes it more adept at handling large datasets with numerous features. While DecisionTreeClassifier can be faster for training and prediction, XGBClassifier provides superior scalability and efficiency in such cases.

TABLE VII. ALGORITHMS COMPARISON

Algorithme Name	Mean	Std
LogisticRegressionCV(LR)	0.839304	019146
XGBClassifier(XGB)	0.831148	012742
KNeighborsClassifier(KNN)	0.838491	023427
DecisionTreeClassifier(DT)	0.845802	012548
RandomForestClassifier(RF)	0.831974	015155
AdaBoostClassifier(ADA)	0.834416	014393
SVC	0.825506	012568

TABLE VI. Parameter Correlation

	Customer ID	Recency	Recency Cluster	Frequency	Frequency Cluster	Monetary	Monetary Cluster	Overall Score	DataF2_Monetary	LTV Cluster	Segment_High-Value	Segment_Low-Value	Segment_Mid-Value
Customer ID	1.00	-0.02	0.02	-0.03	-0.02	-0.09	-0.07	-0.00	-0.06	-0.06	-0.02	0.01	-0.00
Recency	-0.02	1.00	-0.97	-0.29	-0.25	-0.30	-0.19	-0.91	-0.25	-0.24	-0.18	0.79	-0.73
Recency Cluster	0.02	-0.97	1.00	0.28	0.24	0.28	0.17	0.93	0.24	0.23	0.17	-0.83	0.77
Frequency	-0.03	-0.29	0.28	1.00	0.81	0.52	0.39	0.52	0.43	0.41	0.55	-0.34	0.15
Frequency Cluster	-0.02	-0.25	0.24	0.81	1.00	0.49	0.37	0.54	0.42	0.41	0.52	-0.36	0.18
Monetary	-0.09	-0.30	0.28	0.52	0.49	1.00	0.78	0.50	0.68	0.58	0.55	-0.34	0.14
Monetary Cluster	-0.07	-0.19	0.17	0.39	0.37	0.78	1.00	0.42	0.61	0.51	0.73	-0.23	-0.02
Overall Score	-0.00	-0.91	0.93	0.52	0.54	0.50	0.42	1.00	0.42	0.39	0.41	-0.83	0.69
DataF2_Monetary	-0.06	-0.25	0.24	0.43	0.42	0.68	0.61	0.42	1.00	0.86	0.52	-0.28	0.10
LTV Cluster	-0.06	-0.24	0.23	0.41	0.41	0.58	0.51	0.39	0.86	1.00	0.45	-0.26	0.11
Segment_High-Value	-0.02	-0.18	0.17	0.55	0.52	0.55	0.73	0.41	0.52	0.45	1.00	-0.20	-0.16
Segment_Low-Value	0.01	0.79	-0.83	-0.34	-0.36	-0.34	-0.23	-0.83	-0.28	-0.26	-0.20	1.00	-0.94
Segment_Mid-Value	-0.00	-0.73	0.77	0.15	0.18	0.14	-0.02	0.69	0.10	0.11	-0.16	-0.94	1.00

G. Build and Run the ML XGB Model

The “clvcluster” represents the predicted customer clusters based on the features in our dataset. The precision, recall, and f1-score metrics provide insights into how well the XGBClassifier is performing in predicting customer clusters. Based on the provided metrics for each class, we can evaluate the model’s performance for customer segmentation. Higher precision, recall, and f1-scores for a particular cluster indicate that the model is more accurate in predicting customers belonging to that cluster.

The XGB classifier achieved an accuracy of 96% on the training set and an accuracy of 84% on the test set. This suggests that the model has learned the patterns in the training data well and is performing reasonably well on unseen data.

In terms of class-wise metrics:
Class 0:

Precision: 0.89 Recall: 0.94 F1-score: 0.91 Support: 1017
Class 1:

Precision: 0.45 Recall: 0.33 F1-score: 0.38 Support: 184
Class 2:

Precision: 0.56 Recall: 0.32 F1-score: 0.41 Support: 31

These metrics provide insights into the performance of the XGB classifier for each customer cluster. Class 0 has relatively high precision, recall, and F1-score, indicating good predictive performance for this cluster. Class 1 has a lower precision, recall, and F1-score, suggesting that the model struggles more to accurately predict instances in this cluster. Class 2 also has relatively lower precision, recall, and F1-score, indicating room for improvement in predicting instances for this cluster (Table VIII).

TABLE VIII. Precision

Class	Precision	Recall	F1-score	Support
0	0.89	0.94	0.91	1017
1	0.45	0.33	0.38	184
2	0.56	0.32	0.41	31

To further analyze and improve the customer segmentation using the XGBClassifier, we should consider techniques such as hyperparameter tuning and do more feature engineering.

H. Improve the Model

The XGB classifier achieved an accuracy of 93% on the training set and maintained the same accuracy of 84% on the test set. This indicates that the model is still performing well on the test data and is not overfitting, as the training and test accuracies are relatively close.
Let’s look at the class-wise metrics:
Class 0:

Precision: 0.89 Recall: 0.95 F1-score: 0.92 Support: 1017
Class 1:

Precision: 0.49 Recall: 0.36 F1-score: 0.42 Support: 184
Class 2:

Precision: 0.73 Recall: 0.35 F1-score: 0.48 Support: 31

Comparing these metrics (Table VII) with the previous results (Table IX), we can observe some changes. The precision, recall, and F1-scores for class 0 remain relatively similar, indicating that the model’s performance for this cluster is consistent.

For class 1, there is a slight improvement in precision, recall, and F1-score, suggesting that the adjustment to max_depth=4 might have helped the model better capture patterns for this cluster.

Class 2 shows a significant improvement in precision, recall, and F1-score. The model’s ability to predict instances in this cluster has notably improved.

The adjustment in max_depth seems to have improved the model’s performance for some classes while maintaining a similar level of accuracy. However, it’s important to note that further evaluation and analysis are needed to fully assess the effectiveness of the model, such as considering other evaluation metrics and potentially exploring additional model adjustments or techniques without forgetting the specific goals and requirements for customer segmentation.

TABLE IX. ENHANCE MODEL

Class	Precision	Recall	F1-score	Support
0	0.89	0.95	0.92	1017
1	0.49	0.36	0.42	184
2	0.73	0.35	0.48	31

I. False Positive Rate

The receiver operating characteristic (ROC) curve is a graphical representation of the performance of a classification model. It illustrates the relationship between the true positive rate (sensitivity) and the false positive rate (specificity) for different threshold values.

In our case (Fig. 3), we have three classes: segment_low_value, segment_mid_value, and segment_high_value. Each class has its own ROC curve with its corresponding area under the curve (AUC) value.

- ROC of segment_low_value: The AUC value for this class is 0.84. This indicates that the model performs well in distinguishing between the low-value segment and the other classes. The higher the AUC value, the better the model's ability to correctly classify instances of the low-value segment.
- ROC of segment_mid_value: The AUC value for this class is 0.80. This suggests that the model's performance in distinguishing between the mid-value segment and the other classes is slightly lower compared to the low-value segment. However, an AUC of 0.80 still indicates a reasonably good classification performance.
- ROC of segment_high_value: The AUC value for this class is 0.97. This suggests that the model excels in distinguishing between the high-value segment and the other classes. An AUC of 0.97 indicates a high level of accuracy in correctly classifying instances of the high-value segment.

Additionally, we have two overall performance measures:

- Micro-average ROC curve: The AUC value for the micro-average ROC curve is 0.96. This measure takes into account the performance across all classes and provides an aggregated evaluation of the model's overall classification performance. An AUC of 0.96 suggests a high level of accuracy in predicting the correct class across all segments.
- Macro-average ROC curve: The AUC value for the macro-average ROC curve is 0.87. This measure calculates the average AUC value across all classes, giving equal weight to each class. An AUC of 0.87 indicates a good overall performance of the model in distinguishing between the different segments.

Our model demonstrates strong performance in classifying the low-value, mid-value, and high-value segments individually, as indicated by the respective AUC values. The micro-average ROC curve also indicates high accuracy across all segments, while the macro-average ROC curve provides a balanced evaluation of the model's overall performance.

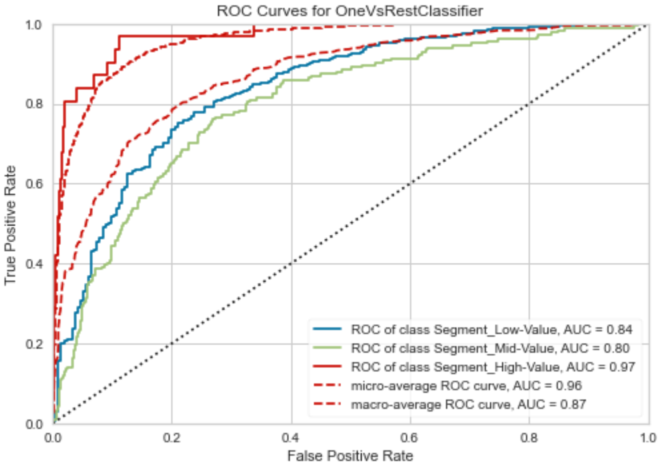


Fig. 3. ROC curve.

J. Precision Recall Curve

The PrecisionRecallCurve shows the tradeoff between a classifier's precision, a measure of result relevancy, and recall, a measure of completeness. For each class, precision is defined as the ratio of true positives to the sum of true and false positives, and recall is the ratio of true positives to the sum of true positives and false negatives.

$$Precision = \frac{TP}{TP + FP} \tag{2}$$

$$Recall = \frac{TP}{TP + FN} \tag{3}$$

where TP denotes true positive, TN stands for true negative, FP means false positive, FN denotes false negative.
A classifier's precision can be thought of as a gauge of its accuracy. It is described for each class as the proportion of true positives to the total of true and false positives. Another way to phrase this question is, "For all instances classified positive, what percent was correct?" The capacity of a classifier to accurately detect all positive cases is measured by recall, which is also known as the completeness of the classifier. It is described as the ratio of true positives to the total of true positives and false negatives for each class. Another way to phrase this question is, for all instances that were actually positive, what percentage was classified correctly? Average precision expresses the precision-recall curve in a single number, which represents the area under the curve. It is determined by computing the weighted average of the precision attained at each threshold, where the weights correspond to the variations in recall between thresholds. When there are class imbalances, the Precision-Recall (PR) curve sheds important light on how well a classification model performs. The average precision value of 0.9 and the Micro-average PR curve for all clusters are both included in our PR curve. The aggregated accuracy and recall for all clusters are shown by the Micro-average PR curve. It offers a comprehensive assessment of the model's capacity to locate favorable occurrences across all classes while taking into account the imbalances in a class distribution (Fig. 4).

The fact that the Micro-average PR curve intersects with the average precision curve at a recall value of 0.7 indicates a crucial point of trade-off in the classification performance. At this threshold, the precision achieved by the model is equal to the average precision of 0.9. It suggests that, on average, the model can correctly identify 90% of positive instances when the recall is 0.7. This threshold represents a balance between precision and recall for the overall classification performance.

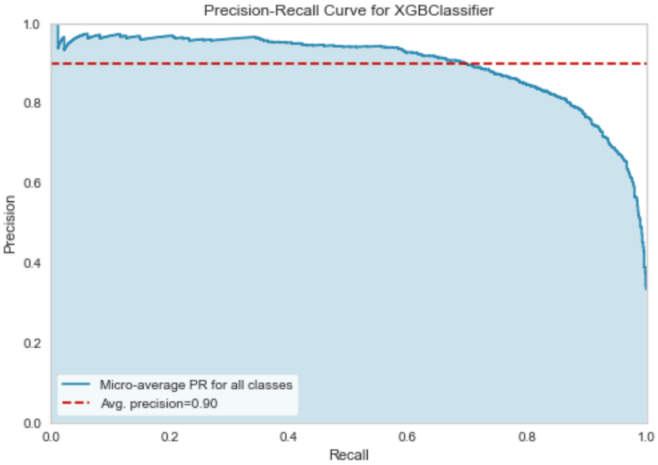


Fig. 4. Precision recall curve.

K. Class Prediction Error

The Yellowbrick ClassPredictionError plot is a twist on other and sometimes more familiar classification model diagnostic tools like the Confusion Matrix and Classification Report [23], [24]. Similar to the classification report, this plot displays a stacked bar chart of the support (number of training samples) for each class in the fitted classification model. As in a Confusion Matrix, each segmented bar displays the percentage of predictions (including false negatives and false positives) for each class. We can utilize a ClassPredictionError to see which classes our classifier is struggling with and, more critically, what false positives it is producing for each class. This frequently enables us to better comprehend the advantages and disadvantages of various models as well as specific difficulties pertaining to your dataset. The class prediction error chart is a fast way to gauge how well the classifier predicts the appropriate classes.
The XGBClassifier demonstrates accurate predictions for the Segment_High_Value class. However, there are instances where it mislabels Segment_Low_Value as Segment_Mid_Value and misclassifies Segment_Mid_Value as Low_Value_Value. In a few cases, it also misclassifies Segment_Mid_Value as Segment_High_Value (Fig. 5).

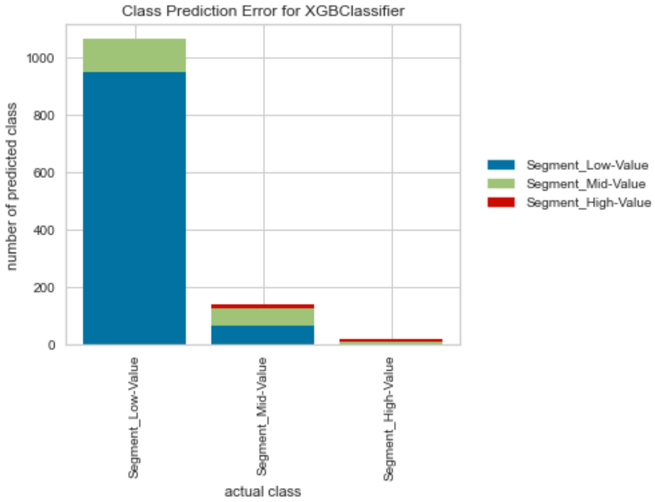


Fig. 5. Class prediction error.

V. Discussion

The work conducted in this study provides valuable insights and tools that can significantly contribute to enhancing customer retention. This work can help improve customer retention by accurately predicting customer segment changes, businesses can identify customers who are at risk of churn or transitioning to lower-value segments. This enables proactive intervention through targeted retention strategies. By offering personalized incentives, tailored communication, and exclusive offers to these customers, businesses can increase their likelihood of staying engaged and loyal. Understanding which customers make up the high-value segment (Segment_High_Value) allows businesses to allocate their resources more effectively. By focusing efforts on retaining these high-value customers, businesses can maximize their return on investment. This can include allocating more advertising budget towards targeted campaigns for high-value customers, providing exceptional customer service, and offering exclusive benefits to strengthen their loyalty. The insights gained from predicting customer segment changes can be used to personalize marketing efforts. By tailoring advertisements, promotions, and communication to specific customer segments, businesses can increase engagement and relevance. This personalized approach enhances the customer experience and strengthens the bond between the customer and the business, leading to improved retention rates. By leveraging predictive models and analytics, businesses can estimate the customer lifetime value (CLV) for different segments. This information helps prioritize efforts and resources towards segments with higher CLV potential By focusing on increasing CLV through customer retention, businesses can optimize their revenue streams and profitability. Anticipating customer segment changes allows businesses to take a proactive approach to customer relationship management. By identifying customers who are likely to transition to higher-value segments, businesses can develop strategies to nurture and guide their journey. This can involve providing personalized recommendations, cross-selling and upselling opportunities, and proactive customer support to enhance their overall experience and increase loyalty.

The work conducted in this study equips businesses with the knowledge and tools to better understand and predict customer segment changes. By leveraging this information, businesses can implement targeted retention strategies, optimize resource allocation, personalize marketing efforts, and proactively manage customer relationships. These efforts collectively contribute to improving customer retention rates and fostering long-term customer loyalty.

Like every research study, our work also faces certain constraints and shortcomings. One limitation is the size of the dataset used for analysis, which may affect the representativeness of the findings. Additionally, the research focused on a specific industry, and the results may not be directly applicable to other sectors.

VI. Conclusions

This paper highlights the effectiveness of predicting customer segment changes to enhance customer retention strategies in the online retail industry. By leveraging machine learning techniques and analytical approaches, businesses can gain valuable insights into customer behavior and forecast future segment transitions. This enables proactive decision-making and targeted actions to retain high-value customers, optimize marketing efforts, and allocate resources efficiently. Future research directions include refining predictive models and algorithms, incorporating external factors, and utilizing real-time data for dynamic segmentation. Advanced customer analytics techniques, like customer journey analysis and sentiment analysis, can provide deeper insights into customer preferences and needs, further enhancing retention strategies. Moreover, predictive analytics can extend beyond customer retention to areas like personalized pricing, inventory management, and supply chain optimization, enabling businesses to deliver a superior customer experience. The study's findings underscore the potential of predicting customer segment changes for enhancing customer retention in the online retail industry. Continued research and innovation in this field will drive the ongoing evolution of customer retention strategies and foster long-term customer loyalty in the competitive online retail landscape.

Authorship Contribution Statement

Lahcen ABIDAR: Conceptualization of this study, Methodology, Software, Data collection, reading and analyzing existing literature, analysis and interpretation of results, Writing - Original draft preparation. Dounia ZAIDOUNI: Participate in the conceptualization of this study, Methodology, analysis and interpretation of results, Review - Original draft preparation. Ikran EL ASRI: Participate in the conceptualization of this study, Methodology, analysis and interpretation of results, Review - Original draft preparation. Abdeslam ENNOUAARY: Supervised the work, participate in the conceptualization of this study, Methodology, Review - Original draft preparation.

References

[1] P. Jana and M. Tiwari, "2 - lean terms in apparel manufacturing," in *Lean Tools in Apparel Manufacturing*, ser. The Textile Institute Book Series, P. Jana and M. Tiwari, Eds. Woodhead Publishing, 2021, pp. 17–45. [Online]. Available: https://www.sciencedirect.com/science/article/pii/B9780128194263000102

[2] T. Parsa Kord Asiabi and R. Tavoli, "A review of different data mining techniques in customer segmentation," *Journal of Advances in Computer Research*, vol. 6, no. 3, 2015.

[3] M. Nilashi, H. Ahmadi, G. Arji, K. O. Alsalem, S. Samad, F. Ghabban, A. O. Alzahrani, A. Ahani, and A. A. Alarood, "Big social data and customer decision making in vegetarian restaurants: A combined machine learning method," *Journal of Retailing and Consumer Services*, vol. 62, 2021.

[4] J. Bauer and D. Jannach, "Improved customer lifetime value prediction with sequence-to-sequence learning and feature-based models," *ACM Transactions on Knowledge Discovery from Data*, vol. 15, no. 5, 2021-06, publisher: Association for Computing Machinery.

[5] Q. Zhang, H. Yamashita, K. Mikawa, and M. Goto, "Analysis of purchase history data based on a new latent class model for RFM analysis," *Industrial Engineering and Management Systems*, vol. 19, no. 2, 2020.

[6] L. Abidar, I. E. Asri, D. Zaidouni, and A. Ennouaary, "A data mining system for enhancing profit growth based on RFM and CLV," in *2022 9th International Conference on Future Internet of Things and Cloud (FiCloud)*, 2022-08, pp. 247–253. [Online]. Available: https://ieeexplore.ieee.org/document/9910557

[7] A. J. Christy, A. Umamakeswari, L. Priyatharsini, and A. Neyaa, "RFM ranking – an effective approach to customer segmentation," *Journal of King Saud University - Computer and Information Sciences*, vol. 33, no. 10, pp. 1251–1257, 2021-12, publisher: King Saud bin Abdulaziz University.

[8] N. P. P. Yuliari, I. K. G. D. Putra, and N. K. D. Rusjayanti, "Customer segmentation through fuzzy c-means and fuzzy RFM method," *Journal of Theoretical and Applied Information Technology*, vol. 78, no. 3, 2015.

[9] R. W. Sembiring Brahmana, F. A. Mohammed, and K. Chairuang, "Customer segmentation based on RFM model using k-means, k-medoids, and DBSCAN methods," *Lontar Komputer : Jurnal Ilmiah Teknologi Informasi*, vol. 11, no. 1, pp. 32–32, 2020-04, publisher: Universitas Udayana.

[10] C. Dullaghan and E. Rozaki, "Integration of machine learning techniques to evaluate dynamic customer segmentation analysis for mobile customers," *International Journal of Data Mining & Knowledge Management Process*, vol. 7, no. 1, 2017.

[11] A. Ahani, M. Nilashi, O. Ibrahim, L. Sanzogni, and S. Weaven, "Market segmentation and travel choice prediction in spa hotels through TripAdvisor's online reviews," *International Journal of Hospitality Management*, vol. 80, 2019.

[12] P. Albuquerque, S. Alfinito, and C. V. Torres, "Support vector clustering for customer segmentation on mobile TV service," *Communications in Statistics: Simulation and Computation*, vol. 44, no. 6, 2015.

[13] M. De Marco, P. Fantozzi, C. Fornaro, L. Laura, and A. Miloso, "Cognitive analytics management of the customer lifetime value: an artificial neural network approach," *Journal of Enterprise Information Management*, vol. 34, no. 2, 2021.

[14] F. Marisa, S. S. S. Ahmad, Z. I. M. Yusoh, T. M. Akhriza, A. L. Maukar, and A. A. Widodo, "Analysis of relationship CLV with 8 core drives using clustering k-means and octalysis gamification framework," *Journal of Theoretical and Applied Information Technology*, vol. 98, no. 20, 2020.

[15] Y. Yuan, K. Dehghanpour, F. Bu, and Z. Wang, "A data-driven customer segmentation strategy based on contribution to system peak demand," *IEEE Transactions on Power Systems*, vol. 35, no. 5, 2020.

[16] A. Mosaddegh, A. Albadvi, M. M. Sepehri, and B. Teimourpour, "Dynamics of customer segments: A predictor of customer lifetime value," *Expert Systems with Applications*, vol. 172, 2021.

[17] K. Khalili-Damghani, F. Abdi, and S. Abolmakarem, "Hybrid soft computing approach based on clustering, rule mining, and decision tree analysis for customer segmentation problem: Real case of customer-centric industries," *Applied Soft Computing Journal*, vol. 73, pp. 816–828, 2018-12, publisher: Elsevier Ltd.

[18] R. Heldt, C. S. Silveira, and F. B. Luce, "Predicting customer value per product: From RFM to RFM/p," *Journal of Business Research*, vol. 127, pp. 444–453, 2021-04, publisher: Elsevier Inc.

[19] L. Abidar, D. Zaidouni, and A. Ennouaary, "Customer segmentation with machine learning: New strategy for targeted actions," in *ACM International Conference Proceeding Series*, 2020.

[20] C. Fan, M. Chen, X. Wang, J. Wang, and B. Huang, "A review on data preprocessing techniques toward efficient and reliable knowledge discovery from building operational data," *Frontiers in Energy Research*, vol. 9, 2021.

[21] D. Jain and S. S. Singh, "Customer lifetime value research in marketing: A review and future directions," *Journal of Interactive Marketing*, vol. 16, no. 2, 2002.

[22] kaggle. onlineretail. [Online]. Available: https://www.kaggle.com/datasets/vijayuv/onlineretail

[23] B. Bengfort and R. Bilbro, “Yellowbrick: Visualizing the Scikit-Learn Model Selection Process,” vol. 4, no. 35, 2019. [Online]. Available: http://joss.theoj.org/papers/10.21105/joss.01075

[24] B. Bengfort, R. Bilbro, N. Danielsen, L. Gray, K. McIntyre, P. Roman, Z. Poh *et al.*, “Yellowbrick,” 2018. [Online]. Available: http://www.scikit-yb.org/en/latest/