WIDS - Predictive Customer Analytics: CLV Forecasting & Behavioral Segmentation Using RFM Methodology

Endterm Report

Karan Bansal (24b3003)

1. Project Objective

Predictive Customer Analytics focuses on using historical customer transaction data to understand purchasing behavior, segment customers, and predict future actions such as churn. These insights enable businesses to improve customer retention, identify high-value customers, and design targeted marketing strategies. The objective of this project is to build an end-to-end predictive analytics system that cleans and preprocesses real-world customer transaction data, performs exploratory data analysis to understand customer behavior, segments customers using the RFM (Recency, Frequency, Monetary) methodology, estimates Customer Lifetime Value (CLV), predicts customer churn using multiple machine learning models, compares and ranks models based on performance, and translates analytical results into actionable business strategies. Overall, the project demonstrates how customer behavioral data can be transformed into meaningful business intelligence that supports strategic decision-making.

2. Dataset Description

The dataset used for this project is the Online Retail dataset (Year 2010 - 2011). It contains transaction-level data from a UK-based online retail company that sells unique all-occasion gifts.

Dataset Characteristics

- Nature: Transactional retail data

- Time period: December 2010 to December 2011
- Granularity: Individual product-level transactions
- Number of records: 541,000
- Number of attributes: 8

Key Variables

Column Name	Description
Invoice	Unique invoice number for each transaction
StockCode	Unique identifier for each product
Description	Product description
Quantity	Number of units purchased
InvoiceDate	Date and time of transaction
Price	Price per unit
Customer ID	Unique customer identifier
Country	Country of the customer

3. Data Cleaning and Preparation

Real-world datasets often contain inconsistencies and noise. Before analysis, several data quality issues were addressed to ensure reliable results.

Some transactions do not contain a Customer ID. Since the objective is customer-level analysis, these records were removed. Retaining such records would prevent accurate aggregation and segmentation.

Transactions with Negative quantity values and Zero or negative prices were excluded from the dataset. These transactions represent product returns, cancellations, or data entry errors and can distort frequency and monetary calculations.

This variable represents the monetary contribution of each transaction and is essential for RFM and churn analysis.

After cleaning, the dataset represents valid purchase behavior and is suitable for customer-level aggregation.

4. Exploratory Data Analysis

Exploratory Data Analysis was conducted to understand overall customer purchasing patterns, geographic distribution, and temporal transaction behavior present in the dataset. The purpose of this analysis was to identify key trends, detect irregularities, and build intuition for subsequent segmentation and predictive modeling tasks. Since the dataset represents real-world retail transactions, EDA plays a critical role in uncovering behavioral insights that cannot be captured through summary statistics alone.

The geographic analysis revealed that the United Kingdom dominates the dataset in terms of transaction volume, indicating that the company primarily operates within the UK market. Countries such as Germany, EIRE, and France follow at a considerable distance, while a long tail of countries contributes relatively fewer transactions. This distribution suggests that the dataset is highly imbalanced across regions, with a strong concentration of customers in the UK and limited representation from other countries. Despite this imbalance, the presence of international customers reflects the company’s expanding global footprint and highlights potential opportunities for region-specific marketing strategies.

Further analysis of average transaction prices across countries showed notable variation in customer spending behavior. Countries such as Singapore, Norway, and Malta exhibited higher average prices compared to other regions, indicating that customers from these markets tend to purchase higher-priced items. However, the distribution of prices across most countries was found to be highly skewed, with a large number of small transactions and a small number of extremely high-value purchases. This skewness is a common characteristic of retail datasets and reflects the presence of

outliers and high-spending customers who contribute disproportionately to overall revenue.

Invoice-level analysis revealed that while most invoices consist of multiple low-value purchases, a few invoices exhibited exceptionally high average prices. Upon closer inspection, these high averages were primarily due to invoices containing only a single high-priced product, rather than bulk purchasing behavior. This highlights the importance of contextual analysis, as raw averages alone can be misleading without examining underlying transaction patterns.

Temporal analysis provided additional insights into seasonal and behavioral trends. The volume of transactions was significantly lower in 2009 compared to 2010 and 2011, which can be attributed to the company's limited geographic presence during its early operational phase. As the company expanded into new countries in subsequent years, transaction volume increased substantially. Monthly trends showed that November experienced the highest number of transactions, which is consistent with festive and holiday shopping periods. Quarterly analysis indicated that the fourth quarter (Q4) accounted for the largest share of transactions, further reinforcing the impact of seasonal demand. Daily patterns revealed higher purchasing activity towards the end of the first week and beginning of the third week of each month, while weekday analysis showed that customers were more active on Thursdays, followed by Tuesdays and Wednesdays.

Product-level analysis revealed that a small number of products account for a large share of total sales volume. Items such as "World War 2 Gliders", "White Hanging Heart", "Assorted Colour Bird", and "Jumbo Bag Red" emerged as the most frequently purchased products. This indicates the presence of strong product preferences and suggests that a limited subset of products drives a significant portion of customer demand.

Overall, the EDA highlights that customer behavior in the dataset is highly heterogeneous, with strong skewness in spending, clear seasonal effects, dominant geographic concentration, and distinct product preferences. These patterns justify the

need for customer segmentation and predictive modeling, as treating all customers uniformly would fail to capture the underlying behavioral diversity present in the data.

5. RFM Analysis

RFM analysis is a widely used customer segmentation technique that evaluates customers based on their purchasing behavior.

RFM Metrics

- Recency (R): Number of days since the customer's most recent purchase
- Frequency (F): Number of invoices generated by the customer
- Monetary (M): Total spending by the customer

Each metric captures a different dimension of customer value.

RFM Calculation

Transactions were aggregated at the customer level. A snapshot date was defined as one day after the last transaction date in the dataset. Recency was calculated as the difference between the snapshot date and the customer's most recent purchase.

RFM Scoring

Each RFM metric was divided into quintiles and assigned scores from 1 to 5:

- Lower recency corresponds to higher scores
- Higher frequency and monetary values correspond to higher scores

The final RFM score is the sum of the individual R, F, and M scores and ranges from 3 to 15.

Customers were segmented based on their RFM scores:

Segment	Description

Best Customers	Recent, frequent, and high spenders
Loyal Customers	Regular customers with consistent purchases
At-Risk Customers	Reduced activity and declining engagement
Lost Customers	Long inactive customers with low engagement

This segmentation clearly differentiates high-value customers from those likely to churn.

6. Customer Lifetime Value

Customer Lifetime Value (CLV) represents the total expected revenue that a business can generate from a customer over the entire duration of their relationship. Unlike short-term performance metrics that focus only on individual transactions, CLV provides a long-term perspective on customer value and enables businesses to prioritize customers based on their future profit potential rather than past activity alone. CLV is a critical concept in customer analytics, as it supports more effective resource allocation, targeted marketing strategies, and long-term retention planning.

In this project, a simplified behavioral approach was used to estimate Customer Lifetime Value using historical transaction data. The CLV was computed as the product of Average Order Value (AOV), purchase Frequency, and estimated Customer Lifespan. Average Order Value was calculated by dividing the total monetary value of purchases by the total number of transactions for each customer, capturing the typical spending per purchase. Purchase Frequency reflects how often a customer interacts with the business, while Customer Lifespan was approximated using the inverse of Recency, which represents the time since the customer’s most recent purchase. This formulation provides an intuitive and interpretable estimate of customer value by combining how much a customer spends, how often they purchase, and how long they are expected to remain active.

Although this CLV model relies on simplified assumptions and does not incorporate discounting or probabilistic lifetime modeling, it remains highly effective for practical business analysis and is widely used in applied customer analytics. The estimated CLV values were further used to segment customers into quartiles, allowing the identification of low-value, medium-value, and high-value customer groups. This segmentation revealed that a relatively small proportion of customers contributes a disproportionately large share of the overall customer value, which is consistent with common observations in retail and e-commerce environments. These insights enable businesses to focus retention and engagement efforts on high-value customers while designing cost-effective strategies for lower-value segments.

Overall, the CLV analysis complements RFM segmentation by adding a forward-looking dimension to customer evaluation. While RFM focuses on historical behavior, CLV provides an estimate of future revenue potential, allowing businesses to move beyond descriptive analytics and towards predictive and prescriptive decision-making.

7. Churn Definition and Modeling

Customer churn refers to the phenomenon where customers discontinue their relationship with a business. In a retail setting, churn typically manifests as prolonged inactivity, where customers stop making purchases over a significant period. Predicting churn enables businesses to proactively identify at-risk customers and design targeted retention strategies.

Customer churn was defined using a time-based heuristic: Customers who have not made a purchase in the last 90 days were labeled as churned, while Customers active within the last 90 days were considered retained.

To predict churn, features derived from RFM analysis were used. These features capture different dimensions of customer behavior and are well-established predictors of churn.

1. Recency measures the number of days since a customer's most recent purchase. It reflects how recently a customer has interacted with the business. Customers with higher recency values are generally more likely to churn, as prolonged inactivity indicates reduced engagement.
2. Frequency represents the total number of purchases made by a customer. Higher frequency often corresponds to stronger customer loyalty and habitual purchasing behavior, making such customers less prone to churn.
3. Monetary value captures the total spending by a customer over the observation period. Customers with higher monetary contributions are typically more valuable and more invested in the business, reducing their likelihood of churn.

Multiple machine learning models were trained and evaluated for churn prediction in order to compare their performance and assess the effectiveness of different modeling approaches. These included Logistic Regression, Random Forest, Decision Tree, K-Nearest Neighbors (KNN), and Gradient Boosting. Logistic Regression was used as the baseline model due to its simplicity and interpretability, while the remaining models were included to capture potential non-linear relationships in customer behavior.

8. Model Comparison and Ranking

MODEL RANKING (by AUC)

	Model	AUC
0	Gradient Boosting	0.773986
1	Logistic Regression	0.772862
2	KNN	0.705915
3	Random Forest	0.697103
4	Decision Tree	0.571600

All models were trained on the same feature set derived from RFM analysis and evaluated on a held-out test set using standard performance metrics, with the Area Under the ROC Curve (AUC) used as the primary criterion for model comparison. AUC was selected as it provides a robust measure of a model's ability to discriminate between churned and retained customers across different classification thresholds.

The results indicate that ensemble-based models generally outperform simpler models. Gradient Boosting achieved the highest AUC among all evaluated models, demonstrating superior performance in capturing complex non-linear patterns in customer behavior. Random Forest also performed competitively, achieving strong predictive performance and higher recall for churned customers compared to baseline models. Logistic Regression, while more interpretable and computationally efficient, showed moderate performance, reflecting the limitations of linear models in capturing complex customer dynamics. Decision Tree and KNN exhibited relatively lower performance, suggesting that these models are less effective for this particular churn prediction task.

Based on the ranking of models using AUC, Gradient Boosting emerged as the best-performing model, followed by Random Forest and Logistic Regression. This ranking highlights the advantage of ensemble learning techniques in predictive customer analytics and demonstrates the importance of model comparison when designing churn prediction systems. The analysis confirms that while simpler models offer interpretability, more advanced ensemble models provide improved predictive accuracy and are better suited for real-world customer behavior modeling.

9. Business Segmentation and Strategy

To translate analytical insights into actionable business decisions, customers were further segmented using a combined framework based on Customer Lifetime Value (CLV) and predicted churn risk. This segmentation approach enables the identification of customer groups not only in terms of their economic value but also their likelihood of disengagement, thereby supporting more informed and strategic marketing decisions. Based on this framework, customers were classified into four categories: VIP, Rescue, Regular, and Low Priority.

VIP customers are those with high CLV and low churn risk, representing loyal and highly valuable customers who contribute significantly to overall revenue. These customers should be targeted with loyalty programs, personalized offers, and premium services in

order to maintain long-term engagement and strengthen customer relationships. Rescue customers are high CLV customers who are predicted to be at high risk of churn and therefore represent the most critical segment for retention efforts. Targeted retention campaigns, discounts, and proactive engagement strategies should be prioritized for this group, as retaining these customers yields the highest return on investment. Regular customers are low CLV customers with low churn risk, indicating stable but moderate engagement. These customers can be maintained through low-cost marketing strategies such as email promotions and general engagement campaigns. Low Priority customers are those with both low CLV and high churn risk, representing customers with limited long-term value. For this group, aggressive retention efforts may not be cost-effective, and minimal marketing resources should be allocated.

This business-oriented segmentation demonstrates how predictive analytics can be directly applied to real-world decision-making. By integrating CLV estimation with churn prediction, businesses can move beyond descriptive analysis and implement data-driven customer management strategies that optimize marketing spend, improve customer retention, and maximize long-term profitability.

10. Conclusion

This project presents a complete end-to-end predictive customer analytics system that transforms raw transactional data into actionable business insights. Beginning with data cleaning and exploratory analysis, the study systematically applied RFM-based customer segmentation, Customer Lifetime Value estimation, and machine learning-based churn prediction to understand and model customer behavior. The integration of descriptive, predictive, and prescriptive analytics demonstrates the full analytical lifecycle, from understanding historical behavior to predicting future outcomes and recommending strategic actions.

The results highlight that historical purchasing patterns are highly informative for predicting customer churn and long-term value. Ensemble learning methods, particularly Gradient Boosting and Random Forest, achieved superior predictive performance compared to simpler models, emphasizing the importance of model selection and

comparison in applied data science. Furthermore, the combination of CLV and churn predictions enabled the development of a business-oriented customer segmentation framework, providing clear guidance for marketing and retention strategies.

Overall, this project illustrates how predictive customer analytics can support data-driven decision-making in real-world business contexts. By leveraging customer transaction data, businesses can identify high-value customers, anticipate churn risks, optimize marketing investments, and improve long-term customer relationships. The methodology and findings of this study demonstrate the practical relevance of machine learning and customer analytics in strategic business planning.

11. References

- Python for Data Analysis - YouTube
- Customer Lifetime Value using Python - Statology
- RFM Analysis - CleverTap
- Churn Prediction using Machine Learning - BCIIT Journal
- IEEE Research on Churn Prediction