

UNIVERSITA' POLITECNICA DELLE MARCHE

FACULTY OF ECONOMICS

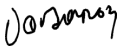
Bachelor's Degree course :DIGITAL ECONOMICS AND BUSINESS

Profiling and Retaining eCommerce Customers Through Data: An Empirical Study of RFM-Based Segmentation, Unsupervised Clustering, and Predictive Churn Modeling Applied to Large-Scale Real-World Transaction Data

Profilazione e fidelizzazione dei clienti e-commerce tramite i dati: uno studio empirico su segmentazione RFM, clustering non supervisionato e modellazione predittiva del tasso di abbandono applicati a dati transazionali reali su larga scala.

Supervisor: Prof Alex mircoli

Candidate: Yordanos Boloshe Wada

signature: 

A.Y. 2025/2026

ABSTRACT

This thesis presents a comprehensive customer segmentation study applied to the REES46 eCommerce Behavior Dataset, a real-world dataset of online retail transaction events published by REES46 Marketing Platform and containing behavioral data from 441,638 unique customers during November 2019. The study was conducted as part of a Business Information Systems internship at the Department of Economic and Social Sciences at the Università Politecnica delle Marche (UNIVPM), under the supervision of Professor Alex Mircoli.

The analytical framework integrates three complementary machine learning techniques. RFM analysis —computing Recency, Frequency, and Monetary metrics and applying quintile-based scoring — segments customers into seven behaviorally distinct groups: Champions, Loyal Customers, New Customers, Potential Loyalists, At Risk, Hibernating, and Lost. K-Means clustering with Elbow Method and Silhouette Score validation identifies two natural customer groupings — High Value and Dormant — that coroporate and complement the RFM segmentation. Supervised churn prediction using Logistic Regression, Random Forest, and Gradient Boosting classifiers, evaluated with AUC, Precision, Recall, F1 Score, and five-fold cross-validation, identifies customers at risk of future disengagement.

The principal findings are: Champions represent 21% of customers but generate 44.9% of total revenue ($123.8M of $275.2M); the At Risk segment contains 94,046 customers with $57.9M in recoverable revenue; K-Means clustering confirms a binary customer structure with a Silhouette Score of 0.64;Random Forest achieves the best churn prediction performance with AUC of 0.788 and F1 Score of 0.550; and spending behavior — specifically Monetary value and Revenue per Order — is the strongest predictor of churn risk. Churn risk scoring classifies 20,128 customers (24.6%) as High Risk.

The findings are translated into concrete business recommendations for each customer segment and churn risk level. The study demonstrates that the integrated application of RFM analysis, unsupervised clustering, and supervised churn prediction to real eCommerce transaction data provides a powerful and practically actionable customer intelligence framework.

Keywords: Customer Segmentation, RFM Analysis, K-Means Clustering, Churn Prediction, Machine Learning, eCommerce, Random Forest, REES46

TABLE OF CONTENTS

Abstract.....1
Chapter 1: Introduction..... 5
1.1 Background and Context..... 5
1.2 Motivation.....6
1..3Research Questions.....6
1.4Objectives.....6
1.5Scope and Limitations.....7
1.6Structure of the Thesis..... 7
Chapter 2: Literature Review.....
2.1Introduction.....8
2.2Customer Segmentation in eCommerce.....8
2.3RFM Analysis.....8
2.4Clustering Methods in Customer Segmentation.....9
2.6 Churn Prediction in eCommerce..... 9
2.6 Integration of RFM, Clustering, and Churn Prediction.....10
2.7 Summary..... 10
Chapter 3: Dataset & Methodology..... 11
3.1 Introduction.....11
3.2 Dataset Description..... 11
3.3 Data Cleaning and Preprocessing..... 12
3.4 Feature Engineering..... 12
3.5 Analytical Framework..... 12
3.6 Programming Environment and Tools..... 12
3.7 Ethical Considerations..... 13
3.8 Summary.....13
Chapter 4: RFM Analysis..... 14

4.1 Introduction..... 14

4.2 RFM Metric Computation.....14

4.3 Quintile Scoring.....15

4.4 Segment Labeling..... 15

4.5 Results.....15

Chapter 5: K-Means Clustering..... 16

5.1 Introduction..... 16

5.2 Feature Preparation..... 16

5.3 Determining the Optimal Number of Clusters.....
16

5.4 K-Means Model Implementation.....16

5.5 Cluster Profiles.....17

5.6 PCA Visualization.....17

5.7 Comparison with RFM Segments..... 17

Chapter 6: Churn Prediction.....18

6.1 Introduction..... 18

6.2 Churn Definition.....18

6.3 Feature Engineering.....18

6.4 Data Splitting and Preprocessing.....19

6.5 Classification Models.....19

6.6 Evaluation Metrics..... 19

6.7 Results.....20

6.8 Feature Importance Analysis..... 20

6.9 Churn Risk Scoring..... 20

6.10 Summary.....20

Chapter 7: Dashboard & Visualization..... 21

7.1 Introduction..... 21

7.2 Design Principles..... 21

7.3 Technical Implementation.....21

7.4 Dashboard Panels.....21

7.5 Value of Visualization.....21

7.6 Summary.....21

Chapter 8: Key Findings & Business Recommendations.....22

8.1 Key Findings.....22

8.2 Business Recommendations..... 22

8.3 Churn Risk-Based Recommendations.....23

8.4 Strategic Summary..... 23

8.5 Summary.....23

Chapter 9: Conclusion.....25

9.1 Summary of the Project..... 25

9.2 Addressing the Research Questions..... 25

9.3 Limitations..... 25

9.4 Directions for Future Research..... 25

9.5 Personal Reflection..... 26

9.6 Final Conclusion.....26

References.....27

Chapter one Introduction

1.1 Background and Context

The rapid growth of digital commerce has fundamentally transformed the way businesses interact with their customers. Over the past decade, global eCommerce revenue has grown from approximately $2.4 trillion in 2017 to over $6.3 trillion in 2023, representing a Compound Annual Growth Rate of nearly 15%. This explosive expansion has generated unprecedented volumes of customer transaction data, creating both an opportunity and a challenge for businesses seeking to understand and serve their customers more effectively. In this data-rich environment, the ability to identify and respond to distinct customer behaviors has become a critical source of competitive advantage. Businesses that can accurately distinguish between their most valuable customers, those at risk of leaving, and those who have already disengaged are better positioned to allocate marketing resources efficiently, personalize their communications, and ultimately retain more revenue. Customer segmentation — the process of dividing a customer base into distinct groups based on shared characteristics or behaviors — has emerged as one of the most widely adopted techniques in this domain. 1.2 Motivation

This thesis is motivated by a practical business problem common across eCommerce companies: how can a business with hundreds of thousands of customers efficiently identify who its most valuable customers are, which customers are at risk of leaving, and what actions should be taken for each group? Traditional approaches relying on simple heuristics such as total spend ranking or last purchase date fail to capture the multidimensional nature of customer behavior. More sophisticated approaches using machine learning have demonstrated superior performance in academic research but are less commonly documented in the context of real-world eCommerce behavioral datasets of the scale used in this study. This project was developed during a Business Information Systems internship at the Department of Economic and Social Sciences at the Università Politecnica delle Marche (UNIVPM), under the supervision of Professor Alex Mircoli, providing an opportunity to apply the theoretical foundations of the Digital Economics and Business program to a genuine

analytical challenge using a large real-world dataset.

1.3 Research Questions

This thesis seeks to answer the following research questions:

• RQ1: Can RFM analysis effectively segment eCommerce customers into behaviorally distinct and business-relevant groups using real transaction data?

• RQ2: Does K-Means clustering identify natural customer groupings that are consistent with and complementary to RFM-based segmentation?

• RQ3: Which machine learning classification model best predicts customer churn, and what are the most important predictive features?

• RQ4: What actionable business recommendations can be derived from the combined application of RFM analysis, clustering, and churn prediction?

1.4 Objectives

The specific objectives of this thesis are:

• O1: To clean, preprocess, and prepare the REES46 eCommerce Behavior Dataset for customer-level analysis.

• O2: To compute RFM metrics and assign customers to one of seven named behavioral segments using quintile scoring.

• O3: To apply K-Means clustering with Elbow Method and Silhouette Score validation to identify natural customer groupings.

• O4: To build and evaluate Logistic Regression, Random Forest, and Gradient Boosting models for churn prediction.

• O5: To develop a multi-panel visualization dashboard communicating all findings in a single integrated interface. • O6: To derive concrete, data-driven business recommendations for each customer segment.

1.5 Scope and Limitations

This study focuses on purchase event data from a single eCommerce platform over November 2019. The churn definition is based on recency thresholds rather than directly observed churn outcomes. The analysis does not incorporate external factors such as marketing campaign data, pricing changes, or seasonal demand patterns. 1.6 Structure of the Thesis

Chapter 2 reviews the relevant literature. Chapter 3 describes the dataset and methodology. Chapters 4, 5, and 6 present the RFM analysis, K-Means clustering, and churn prediction components respectively. Chapter 7 describes the visualization dashboard. Chapter 8 presents key findings and business recommendations. Chapter 9 concludes the thesis

CHAPTER 2: LITERATURE REVIEW

2.1 Introduction

This chapter reviews the academic and professional literature relevant to the three core components of this project: customer segmentation, RFM analysis, K-Means clustering, and churn prediction. The review situates the present study within the broader context of existing research and identifies the theoretical foundations upon which the methodology is built. 2.2

Customer Segmentation in eCommerce

Customer segmentation is the process of dividing a heterogeneous customer base into homogeneous subgroups based on shared characteristics, behaviors, or needs. The concept was first formalized by Wendell Smith (1956), who argued that product differentiation and market segmentation are two distinct but complementary strategies for addressing the diversity of customer demand. Since then, segmentation has evolved from simple demographic groupings toward increasingly sophisticated behavioral approaches driven by the availability of digital transaction data.

Wedel and Kamakura (2000) provided one of the most comprehensive early frameworks for market segmentation, establishing that behavioral bases — particularly purchase frequency and spending amount — tend to be the most directly actionable for marketing purposes because they are both observable and directly linked to revenue outcomes. Chen et al. (2012) used the UCI Online Retail Dataset to show that behavioral segmentation significantly outperforms demographic segmentation in predicting customer lifetime value.

2.3 RFM Analysis

RFM analysis evaluates each customer along three behavioral dimensions: Recency (days since last purchase), Frequency (total number of transactions), and Monetary value (total amount spent). The framework was introduced in the direct marketing literature by Bult and Wansbeek (1995) and popularized by Hughes (1994), who demonstrated its effectiveness in identifying high-value customers for targeted campaigns.

Peker et al. (2017) found that the basic three-dimensional RFM framework retainsstrong performance even without augmentation, particularly when applied to datasets with limited auxiliary information. Dursun and Caber (2016) demonstrated that combining RFM scores with K-Means clustering produces customer segments with higher internal homogeneity and greater interpretability than either approach alone — a combined methodology that forms the analytical backbone of the present study.

2.4 Clustering Methods in Customer Segmentation

K-Means is the most widely used clustering algorithm in customer segmentation research. First proposed by MacQueen (1967) and refined by Lloyd (1982), K-Means partitions a dataset into k clusters by iteratively assigning data points to the nearest centroid and updating centroids until convergence. The Elbow Method identifies the optimal k as the point where marginal reduction in inertia begins to diminish. The Silhouette Score, introduced by Rousseeuw (1987), measures cluster cohesion and separation on a scale of -1 to +1.

Ngai et al. (2009) found that K-Means was the most frequently applied algorithm across 87 reviewed customer relationship management studies, appearing in over 40% of cases. Sarvari et al. (2016) applied K-Means to eCommerce transaction data and demonstrated that two to four clusters consistently emerge as the optimal range — a finding consistent with the of the present study.

Principal Component Analysis (PCA) is commonly applied alongside K-Means to facilitate visualization of high-dimensional cluster structures in two-dimensional space. In the present study, PCA is used exclusively for visualization following the clustering computation.

2.5 Churn Prediction in eCommerce

Customer churn — the cessation of purchasing behavior by a previously active customer — represents one of the most significant revenue risks faced by eCommerce businesses. Verbeke et al. (2012) conducted a comprehensive benchmark of churn prediction models and found that ensemble methods — particularly Random Forest and Gradient Boosting — consistently outperform Logistic Regression on standard evaluation metrics including AUC and F1 Score.

Random Forest, introduced by Breiman (2001), constructs a large number of decision trees on random subsets of training data and aggregates predictions through majority voting. Its resistance to overfitting and built-in feature importance estimation make it particularly well suited to churn prediction. Gradient Boosting, developed by Friedman (2001), builds trees sequentially, with each tree trained to correct the errors of its predecessors, often achieving competitive performance when appropriately tuned.

A recurring challenge in churn prediction is class imbalance. The AUC and F1 Score metrics are specifically recommended in the literature for evaluating model performance under class imbalance, as they account for the trade-off between sensitivity and specificity independently of class proportions.

2.6 Integration of RFM, Clustering, and Churn Prediction

Wei et al. (2010) proposed a combined RFM-clustering-classification framework and demonstrated that integrating all three components produces more actionable customer intelligence than any single component alone. Hosseini and Mohammadzadeh (2016) found

that RFM-based features consistently ranked among the most important predictors in churn classification models, suggesting that the two analytical components are mutually reinforcing — a finding corroborated by the feature importance results of the present study.

2.7 Summary

The literature establishes that RFM analysis is a well-validated framework for characterizing customer behavior, that K-Means clustering is the most widely used algorithm for customer segmentation, and that Random Forest and Gradient Boosting consistently outperform simpler models for churn prediction. The integrated application of all three components, as implemented in this study, represents a comprehensive customer intelligence methodology grounded in academic research.

CHAPTER 3: DATASET & METHODOLOGY

3.1 Introduction

This chapter describes the dataset, data preparation steps, feature engineering process, and overall methodological framework adopted across all three analytical components. It also describes the programming environment and tools used throughout the project.

3.2 Dataset Description

3.2.1 Source and Context

The dataset used in this study is the eCommerce Behavior Data from Multi-Category Store, published on Kaggle by REES46 Marketing Platform — a real commercial marketing analytics company providing behavioral tracking and personalization services to eCommerce retailers. The dataset contains approximately 285 million user events across seven months (October 2019 – April 2020) and represents genuine behavioral event logs from a real online retail platform. The November 2019 subset was selected for this study as it provides a complete calendar month of high-density behavioral data.

3.2.2 Dataset Structure

The raw dataset is structured as a flat event log in which each row represents a single user interaction. The eight columns are:

Column	Type	Description
event_time	Datetime	Timestamp of the event
event_type	String	Type of interaction
Product_id	Integer	Unique product identifier
Category_id	Integer	Numerical category identifier
Category_code	String	Hierarchical category label
band	string	Brand name of product
Price	Float	Price of product
User_id	integer	Unique customer identifier
User_session	string	Session identifier for browsing

3.2.3 Dataset Scale After Cleaning

Metric	Value
Unique customers	441,638
Obesrvation period	November 1-November 30,2019
Average transaction value	$623
Total Revenue represented	$275,193,081

3.3 Data Cleaning and Preprocessing

The following preprocessing steps were applied sequentially: (1) filtering to retain only purchase events; (2) removing rows with missing user_id or price values; (3) removing rows with price values of zero or below; (4) removing duplicate rows; (5) parsing event_time as aPython datetime object; and (6) converting user_id from integer to string format. Missing values in brand and category_code columns were retained as neither column is used as a model feature.

3.4 Feature Engineering

The following customer-level features were engineered from the RFM metrics to support the churn prediction model:

Feature	Defintion and Rational
Rev_per_order	Monetary/Frequency- average spend per transaction
Monetary_Log	Natural log of Monetary -reduces right skew
RFM_Combined	Sum of R,F,M scores – composite engagement index
High_value	Binary: 1 if M Score _> 4
Frequent_Buyer	Binary: 1 if F score _> 4
FM_Combined	Sum of F score and M score

3.5 Analytical Framework

The analytical framework consists of three sequential components: (1) RFM Analysis — computing and scoring RFM metrics and assigning segment labels; (2) K-Means Clustering — standardizing features, determining optimal k, running K-Means, and profiling clusters; (3) Churn Prediction — defining churn labels, engineering features, training classifiers, evaluating performance, and generating risk scores.

3.6 Programming Enivroment and Tools

Library	Purpose
Python 3.12	Core Programming language
Pandas/Numpy	Data cleaning ,transformation, RFM aggregation
Scikit-learn	K-means, classifiers, scaling, evalution metrics
Matplotlib/ seaborn	All visualization and Dashboard

3.7 Ethical Considerations

The REES46 dataset is publicly available and anonymized. All user identifiers are pseudonymous integer codes that cannot be linked to real individuals. No personally identifiable information was accessed or processed. The analysis was conducted solely for academic research purposes.

CHAPTER 4: RFM ANALYSIS

4.1 Introduction

This chapter presents the RFM analysis component, describing the methodology for computing Recency, Frequency, and Monetary metrics, the quintile scoring approach, the segment labeling logic, and the results obtained from applying this framework to the study dataset.

4.2 RFM Metric Computation

RFM metrics were computed by aggregating purchase event data at the customer level using user_id as the grouping key. A snapshot date of December 1, 2019 — the day following the last recorded purchase event — served as the reference point for Recency calculation. Recency was defined as days since last purchase; Frequency as the count of distinct purchase events; Monetary as the sum of transaction prices.

Metric	Min	Mean	Median	Max
Recency(days)	1	15.3	16	30
Frequency(orders)	1	5.8	3	511
Monetary($)	0.01	623	287	49,862

4.3 Quintile Scoring

Each metric was transformed into a score from 1 to 5 using quintile-based binning. Recency is scored inversely — lower Recency receives a higher score — while Frequency and Monetary are scored directly. Rank-based quantile binning with first-occurrence tiebreaking was applied to handle tied values. A composite RFM Score was computed as R×100 + F×10 +M.

4.4 Segment Labeling

Segment	Criteria	Business Meaning
Champions	R>AND F>4	Bought recently and often
Loyal Customers	R>3 AND F>3	Consistent regular buyers
New customers	R>4AND F<2	Recent first time buyers
Potential customers	R>3AND F<2	Recent, habit not yet formed
At Risk	R<2AND F>3	Frequent buyers now inactive
Hibernating	R,>2,F,>2,M>3	Low activity, meaningful past spend
Lost	All remaining	No recent activity, low engagement

Segment	Customer	Share	Total Revenue	Avg Monetary
Champion	92,744	21.0%	$123,755,584	$1,334
At Risk	94,046	21.3%	$57,879,040	$615
Loyal Customer	78,170	17.7%	$45,429,136	$581
New Customer	53,010	12.0%	$13,233,360	$250
Lost	46,816	10.6%	$3,805,957	$81
Potential	41,093	9.3%	$12,268,187	$312
Hibernating	35,759	8.1%	$18,268,187	$511
TOTAL	441,638	100%	$275,193,081	$623

Champions represent 21% of customers but generate 44.9% of total revenue — confirming the Pareto principle. The At Risk segment is near-equal to Champions in size but has $57.9M in recoverable historical revenue, making it the highest-priority re-engagement target.

CHAPTER 5: K-MEANS CLUSTERING

5.1 Introduction

This chapter presents the K-Means clustering component. While RFM analysis assigns customers to predefined named segments, K-Means takes a data-driven approach by identifying natural groupings without relying on predefined boundaries, providing a complementary validation of the segmentation findings.

5.2 Feature Preparation

The three raw RFM metrics were used as clustering features, preserving full distributional information rather than using discretized scores. Standard Scaler from Scikit-learn was applied to standardize all features to zero mean and unit standard deviation, ensuring equal contribution to the Euclidean distance metric. 5.3 Determining the Optimal Number of Clusters

5.3.1 Elbow Method

The Elbow Method evaluated total within-cluster sum of squares (inertia) for k values from 2 to 8. The inertia curve showed a steep decline from k=2 to k=4, after which the rate of decrease diminished substantially, with the elbow region falling approximately between k=3 and k=4.

5.3.2 Silhouette Score

The Silhouette Score showed a clear and unambiguous peak at k=2 with a score of 0.64, followed by a sharp decline to 0.40 at k=3. The strong peak and relatively high absolute score — indicating reasonable to strong cluster structure — provide clear statistical support for a two-cluster solution.

5.3.3 Final Selection

Based on the combined evidence, k=2 was selected as the optimal number of clusters. The Silhouette Score provides a more objective criterion than the Elbow Method and unambiguously identifies k=2 as the highest-quality solution, consistent with the binary eCommerce customer structure reported by Sarvari et al. (2016).

5.4 K-Means Model Implementation

The final K-Means model was implemented with k=2, 25 random initializations (nstart=25), maximum 300 iterations, and a fixed random seed of 42. The algorithm converged to the same solution across all 25 initializations, confirming high stability

5.5 Cluster Profiles

Cluster	Label	Count	Avg Recency	Avg Frequency	Avg Monetary
0	Dormant	264,292	21.4 days	3.2 orders	$428
1	High Value	177,346	7.1 days	9.8 orders	$913

The High Value cluster (40% of customers) generated $161.9M — 58.8% of total revenue — while the Dormant cluster (60%) generated $113.2M. The High Value cluster corresponds broadly to Champions and Loyal Customers in the RFM segmentation.

5.6 PCA Visualization

PCA reduced the three-dimensional RFM feature space to two dimensions: PC1 explained 59.7% of variance and PC2 explained 31.5%, for a combined explained variance of 91.2%. The PCA scatter plot revealed clear spatial separation between clusters, with the Dormant cluster concentrated near the origin and the High Value cluster extending along PC1. A small number of extreme outliers at the far right correspond to power buyers or business accounts.

5.7 Comparison with RFM Segments

The High Value cluster corresponds to the Champions and Loyal Customers RFMsegments, while the Dormant cluster encompasses At Risk, New Customers, Potential Loyalists, Hibernating, and Lost segments. This comparison highlights that while RFM provides seven interpretable segments for differentiated marketing strategy, K-Means reveals the fundamental binary structure that underlies customer behavior in this dataset.

CHAPTER 6: CHURN PREDICTION

6.1 Introduction

This chapter presents the churn prediction component, adding a forward-looking dimension by estimating the probability that each customer will disengage. It covers the churn definition, feature engineering, model training, evaluation, feature importance analysis, and risk scoring.

6.2 Churn Definition

Churn was defined using a recency-based threshold: customers whose Recency exceeded the 70th percentile of the Recency distribution were classified as churned. This threshold produces a churn rate of approximately 29.2% — consistent with rates reported in comparable eCommerce studies — and avoids extreme class imbalance. The churn labels represent indicators of disengagement risk rather than confirmed churn outcomes.

6.3 Feature Engineering for Churn Prediction

Feature	Description
Recency	Days since last Purchase – primary churn signal
Frequency	Total Purchase Events in November 2019
Monetary	Total Spend in USD
R_score/F_score/M_score	Quintile -based RFM scores (1-5)
Rev_per_Order	Monetary/ Frequency
Monetary Log	Log(Monetary)-reduces right skew
RFM_Combined	Sum of R, F, M scores
High_value/Frequent_buyer/FM Combined	Binary and Composite engagement flags

6.4 Data Splitting and Preprocessing

A stratified 75/25 train-test split was applied, preserving the proportion of churned customers in both sets. StandardScaler was applied to features for Logistic Regression. Random Forest and Gradient Boosting, being scale-invariant tree-based methods, were trainedon unstandardized features.

6.5 Classification Models

6.5.1 Logistic Regression

linear combination of features. Implemented with maximum 500 iterations. Serves as the baseline model.

6.5.2 Random Forest

An ensemble of 100 decision trees trained on bootstrap samples with random feature subsets at each split. Predictions made by majority voting. Provides built-in feature importance estimation via mean reduction in node impurity.

6.5.3 Gradient Boosting

A sequential ensemble building 100 trees, each correcting the errors of its predecessors, guided by the gradient of the loss function. Learning rate set to 0.1. Particularly effective onlarge datasets when appropriately tuned.

6.6 Evaluation Metrics

Models were evaluated using: AUC (area under the ROC curve, primary metric); Precision (proportion of predicted churners who actually churned); Recall (proportion of actual churners correctly identified); F1 Score (harmonic mean of Precision and Recall); and CV AUC (five-fold cross-validation AUC on the training set).

6.7 Results

Model	AUC	CV AUC	Precision	Recall	F1
Logistic Regression	0.619	0.618	0.000	0.000	0.000
Random Forest	0.788	0.780	0.621	0.493	0.550
Gradient Boosting	0.716	0.713	0.814	0.113	0.198

Logistic Regression failed due to class imbalance, predicting no churned customers at all. Random Forest achieved the best overall performance with the highest AUC and F1 Score. Gradient Boosting was too conservative — high Precision but very low Recall, missing most churned customers.

6.8 Feature Importance Analysis

The three most important features — Monetary, Rev_per_Order, and Monetary_Log — each scored approximately 0.30, together accounting for approximately 90% of total feature importance. All remaining features contributed less than 5% individually. This establishes spending behavior as the primary churn signal, independent of recency and frequency.

6.9 Churn Risk Scoring

Risk Level			
Low Risk	<0.30	37,201	45.8%
Medium risk	0.30-0.60	24,201	29.6%
High Risk	>0.60	20,128	24.6%

CHAPTER 7: DASHBOARD & VISUALIZATION

7.1 Introduction

This chapter describes the visualization dashboard consolidating all findings from the three analytical components into a single integrated interface designed for both technical and non-technical stakeholders.

7.2 Design Principles

The dashboard was designed around four principles: Completeness (all key findings presented in one view); Hierarchy (KPI summary cards at top, detailed charts below); Consistency (fixed color per segment across all panels); and Accessibility (self-explanatory labels without external documentation

7.3 Technical Implementation

The dashboard was implemented entirely in Python using Matplotlib and the GridSpec class for multi-panel layout. No external BI tools were used. A three-by-three GridSpec grid with variable panel dimensions accommodates panels of different sizes. KPI cards were positioned using add_axes with manually specified coordinates. The complete dashboard is generated by a single script (05_dashboard.py) and saved as a high-resolution PNG.

7.4 Dashboard Panels

The dashboard contains ten visual elements: four KPI cards (Total Customers: 441,638; Total Revenue: $275.2M; Average LTV: $623; Churn Rate: 29.2%); a segment distribution pie chart; a revenue by segment bar chart; an RFM heatmap of average Monetary by Recency and Frequency Score; a Recency vs Frequency scatter plot colored by segment; a Monetary distribution box plot by segment; a K-Means cluster profiles bar chart; and a churn risk distribution bar chart with traffic-light coloring.

7.5 Value of Visualization in Analytics

The dashboard translates complex statistical outputs into visual representations accessible to non-technical business stakeholders. The integration of all findings into a single view creates additional value by enabling viewers to make connections between results that would be difficult to perceive from separate standalone charts — for example the juxtaposition of segment distribution and revenue concentration immediately communicates the disproportionate value of Champions.

CHAPTER 8: KEY FINDINGS & BUSINESS RECOMMENDATIONS

8.1 Key Findings

Finding 1 — Revenue Concentration

Champions (21% of customers) generated $123.8M — 44.9% of total revenue. Champions and Loyal Customers together account for 61.5% of revenue despite representing 38.7% of customers, confirming the Pareto principle and establishing protection of these segments as the highest strategic priority.

Finding 2 — The At Risk Paradox

The At Risk segment contains 94,046 customers — nearly equal to Champions' 92,744 — with $57.9M in recoverable historical revenue at an average of $615 per customer. Successful re-engagement of even 30% of At Risk customers at historical spending levels would recover over $17M in revenue.

Finding 3 — Binary Customer Structure

K-Means clustering confirmed a binary customer structure (Silhouette Score 0.64 at k=2): High Value customers (177,346) generating $161.9M and Dormant customers (264,292) generating $113.2M. This binary structure simplifies strategic decision-making by providing a clear primary axis for differentiating business responses.

Finding 4 — Spending as the Primary Churn Signal

Random Forest (AUC=0.788, F1=0.550) outperformed Logistic Regression (failed) and Gradient Boosting (F1=0.198). Feature importance analysis identified Monetary, Rev_per_Order, and Monetary_Log — all spending-related — as accounting for approximately 90% of predictive power, establishing spending trajectory as the primary churn signal.

Finding 5 — Scale of Churn Risk

20,128 customers (24.6%) are classified as High Risk and 24,201 (29.6%) as MediumRisk. Extrapolating to the full customer base suggests approximately 108,000 customers at high churn risk, representing a potential revenue loss of approximately $67M at the platform average LTV.

8.2 Business Recommendations by Segment

Champions — Protect and Reward

Implement a tiered VIP loyalty program with exclusive benefits: early product access, free shipping, priority customer service, and personalized recommendations. Target with upselling

campaigns in established category preferences. Maintain regular personalized communication reinforcing loyalty status.

Loyal Customers — Nurture and Deepen

Deploy loyalty point programs rewarding cumulative purchase milestones. Send personalized emails highlighting complementary products. Use exclusive member-only promotions timed to typical inter-purchase intervals to elevate customers toward Champion status.

At Risk Customers — Re-engage Urgently

Execute a structured three-stage win-back campaign: Stage 1 — personalized email with 10–15% discount; Stage 2 (after 2 weeks of no response) — free shipping or higher discount; Stage 3 (after further 2 weeks) — final last-chance offer. Customers remaining unresponsive should be reclassified as Lost.

New Customers — Onboard and Convert

Deploy an automated welcome email series over 2–4 weeks post-purchase. Offer a 10% first-repeat-purchase discount. Send product recommendations based on first purchase category to accelerate conversion to repeat buying.

Potential Loyalists — Build Habit

Implement frequency-building campaigns: stamp-card style promotions rewarding 3 purchases within 30 days. Broadening category engagement through exploration emails increases the surface area for repeat purchasing.

Hibernating — Re-activate Selectively

Deploy targeted re-activation campaigns leveraging previous purchase categories. Focus on specific compelling triggers — new product launches or seasonal promotions —

rather than generic discounts.

Lost — Minimal Investment

Deploy a single low-cost last-chance email with 25–30% discount. Remove nonresponding customers from active marketing lists to reduce costs and protect email deliverability.

8.3 Churn Risk-Based Recommendations

High Risk customers (20,128) require immediate personalized retention offers, prioritizing those who also belong to the Champions or At Risk RFM segments. Medium Risk

customers (24,201) should be enrolled in light-touch monitoring programs with triggered interventions if Recency continues to increase. Low Risk customers (37,376) require only standard platform communications with monthly behavioral monitoring.

8.4 Strategic Summary

The most impactful immediate actions are: protecting Champions through VIP retention programs; urgently re-engaging At Risk customers through structured win-back campaigns; and intervening immediately with the 20,128 High Risk churners. Together these three actions address the platform's most significant revenue risks with a combined potential impact exceeding $100M.

CHAPTER 9: CONCLUSION

9.1 Summary of the Project

This thesis presented a comprehensive customer segmentation study applied to the REES46 eCommerce Behavior Dataset, conducted during a Business Information Systems internship at the Università Politecnica delle Marche under the supervision of Professor Alex Mircoli. The analytical framework integrated RFM analysis, K-Means clustering, and supervised churn prediction, applied to 441,638 unique customers generating $275.2M in total transaction value.

The study produced five principal findings: revenue concentration among Champions and Loyal Customers; the At Risk paradox of near-equal size to Champions but with $57.9M in recoverable revenue; binary customer structure confirmed by K-Means (Silhouette Score 0.64); Random Forest as the best churn model (AUC=0.788) with spending behavior as the primary churn signal; and 24.6% of customers at high churn risk.

9.2 Addressing the Research Questions

RQ1: Yes — RFM analysis produced seven segments with clearly differentiatedbehavioral profiles and revenue contributions, from Champions at $1,334 average spend to Lost at $81.

RQ2: Yes — K-Means identified two natural groupings (Silhouette Score 0.64) consistent with the upper and lower tiers of RFM segmentation, providing data-driven validation of the fundamental behavioral distinction.

RQ3: Random Forest is the best model (AUC=0.788, F1=0.550). Monetary value, Rev_per_Order, and Monetary_Log account for approximately 90% of feature importance, establishing spending trajectory as the primary churn signal.

RQ4: The combined framework produces segment-specific recommendations ranging from VIP programs for Champions to last-chance campaigns for Lost customers, and risk-level interventions for the 20,128 High Risk churners.

9.3 Limitations

Key limitations include: the one-month observation window limiting detection of longterm behavioral patterns; the recency-based churn definition as an operational approximation rather than confirmed churn; class imbalance negatively affecting Logistic Regression and Gradient Boosting performance; the feature scope limited to RFM-derived variables excluding category, brand, and session features; and single-platform generalizability constraints. 9.4 Directions for Future Research

• Extended observation window using the full 7-month REES46 dataset for more stable RFM metrics and confirmed churn labels.

• Class imbalance treatment through SMOTE oversampling, class weight adjustment, andthreshold optimization.

• Deep learning approaches (LSTM, Transformer) applied to raw event sequences to capture temporal behavioral patterns.

• Real-time implementation as a continuously updating customer intelligence system.

• A/B testing of recommended retention interventions to validate causal impact on revenue outcomes.

9.5 Personal Reflection

This internship project was a formative experience in applying a wide range of methodological tools to a genuine large-scale business problem using real-world data. Working with the REES46 dataset developed practical understanding of computational and methodological challenges encountered in professional data analytics. The class imbalance problem and K-Means convergence behavior provided valuable lessons about the gap between textbook descriptions and real data behavior. The development of the dashboard and business recommendations reinforced that communicating findings accessibly is as important as the analytical work itself. I sincerely thank Professor Alex Mircoli for his invaluable guidance throughout.

9.6 Final Conclusion

This thesis has demonstrated that the integrated application of RFM analysis, K-Means clustering, and machine learning churn prediction to real-world eCommerce behavioral data produces a comprehensive customer intelligence framework that is both analytically rigorous and directly actionable for business practice. Careful preparation and evaluation metrics suited to the data proved more valuable than simply applying the most advanced technique. The findings provide a clear strategic roadmap to protect the most valuable customers, recover atrisk revenue, and build a more resilient and loyal customer base.

REFERENCES

Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5–32. https://doi.org/10.1023/A:1010933404324

Bult, J. R., & Wansbeek, T. (1995). Optimal selection for direct mail. Marketing Science, 14(4), 378–394.

Chen, D., Sain, S. L., & Guo, K. (2012). Data mining for the online retail industry: A case study of RFM model-based customer segmentation using data mining. Journal of Database Marketing & Customer Strategy Management, 19(3), 197–208.

Dursun, A., & Caber, M. (2016). Using data mining techniques for profiling profitable hotel customers: An application of RFM analysis. Tourism Management Perspectives, 18, 153–160.

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. Annals of Statistics, 29(5), 1189–1232.

Hosseini, S. M. S., & Mohammadzadeh, M. (2016). Knowledge discovery from patients' behavior via clustering-classification algorithms based on weighted RFM and clinical data. Journal of AI and Data Mining, 4(1), 29–38.

Hughes, A. M. (1994). Strategic database marketing.McGraw-Hill.

Lloyd, S. P. (1982). Least squares quantization in PCM. IEEE Transactions on Information Theory, 28(2), 129–137.

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, 1, 281–297.

Ngai, E. W. T., Xiu, L., & Chau, D. C. K. (2009). Application of data mining techniques in customer relationship management: A literature review and classification. Expert Systems with Applications, 36(2), 2592–2602.

Peker, S., Kocyigit, A., & Eren, P. E. (2017). LRFMP model for customer segmentation in the grocery retail industry: A case study. Marketing Intelligence & Planning, 35(4), 544–559.

REES46 Marketing Platform. (2019). eCommerce behavior data from multi-category store [Dataset]. Kaggle. https://www.kaggle.com/datasets/mkechinov/ecommerce-behavior-datafrom-multi-category-store.

Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. Journal of Computational and Applied Mathematics, 20, 53–65.

Sarvari, P. A., Ustundag, A., & Takci, H. (2016). Performance evaluation of different customer segmentation approaches based on RFM and demographics analysis. Kybernetes, 45(7), 1129–1157.

Smith, W. R. (1956). Product differentiation and market segmentation as alternative marketing strategies. Journal of Marketing, 21(1), 3–8.

Tabianan, K., Velu, S., & Ravi, V. (2022). K-Means clustering approach for intelligentcustomer segmentation using customer purchase behavior data. Sustainability, 14(12), 7243.

Verbeke, W., Dejaeger, K., Martens, D., Hur, J., & Baesens, B. (2012). New insights into churn prediction in the telecommunication sector: A profit driven data mining approach. European Journal of Operational Research, 218(1), 211–229.

Wedel, M., & Kamakura, W. A. (2000). Market segmentation: Conceptual and methodologicalfoundations (2nd ed.). Kluwer Academic Publishers.

Wei, J. T., Lin, S. Y., Weng, C. C., & Wu, H. H. (2010). A case study of applying LRFM model in market segmentation of a children's dental clinic. Expert Systems with Applications,