

**OPTIMIZING HOME DELIVERY IN SMALL GROCERY STORES: A Data-Driven
Strategy for Customer Segmentation and Marketing**

Manideep Racharla (M.S, Analytics)

School of Professional Studies

Saint Louis University

AA-5962 – Applied Analytics Master’s Project-3

Instructor: Dr. Ravindranath Arunasalam

1. OPTIMIZING HOME DELIVERY IN SMALL GROCERY STORES

1.1.ABSTRACT

This paper focuses on a critical examination of customer segmentation and targeted marketing strategies in the domain of small grocery stores providing home delivery services. Drawing upon machine learning classification algorithms that leverage Recency, Frequency, Monetary, and Time (RFM) data (Ullah, 2023), this research endeavors to distill actionable insights from customer behaviors to foster personalized marketing. Complementing this, a K-Means clustering approach is employed to intelligently categorize customers based on purchase behavior data, aiming to enhance the precision of marketing initiatives tailored to individual consumer segments.

The main aim of the project is to harness these advanced analytical techniques to empower small grocery stores in optimizing their marketing strategies, thus catalyzing growth in their home delivery services. The significance of the study is anchored in its potential to offer a granular, data-driven understanding of customer dynamics, contributing to the competitive agility of small retailers in the fast-evolving e-commerce landscape. By merging theoretical frameworks with empirical data, the research intends to provide a comprehensive resource for small grocery stores to navigate and thrive within the complexities of modern consumer marketing.

Keywords: Customer Segmentation, Targeted Marketing, Small Grocery Stores, Churn, Consumer Behavior, Recency, frequency, Monetary (RFM) Analysis, K-Means Clustering, Direct delivery.

1.2.Context & Motivation

In the evolving landscape of retail, particularly within the grocery sector, small businesses face increasing competition not only from large supermarkets but also from online platforms. This shift necessitates innovative approaches to customer engagement and retention, where personalized marketing and efficient service delivery become crucial. Small grocery stores often lack access to advanced analytical tools and strategies that larger corporations utilize, leaving them at a disadvantage in this highly competitive environment.

My motivation for this study stems from personal experience. Growing up in a family that ran a small grocery store, I witnessed firsthand the challenges faced by small businesses in retaining customers and staying competitive. I saw how the success of the store often relied on loyal customers and the ability to adapt to their needs. However, the lack of data-driven decision-making tools made it difficult to implement targeted marketing or re-engagement strategies effectively.

This personal connection inspired me to explore how small grocery stores can leverage affordable and feasible analytics solutions to address these challenges. By focusing on customer segmentation and predictive modeling, this study aims to provide actionable insights and practical tools that empower small businesses to compete in the modern retail environment.

1.3. Research Questions:

1. *How can customer segmentation improve targeted marketing?*
2. *How can churn prediction help small grocery stores?*
3. *How can RFM analysis be used to segment customers?*
4. *How accurately can machine learning models predict churn?*
5. *What insights can be derived from clustering, and how can they improve marketing strategies?*
6. *What marketing strategies can be implemented to improve retention rates of customers in Cluster 3??*

1.4. Aims & Objectives

This research aims to validate the effectiveness of using RFM analysis and K-Means clustering for real-time customer data processing, facilitating enhanced decision-making in marketing and service delivery for small grocery stores.

1.5. Thesis Overview

The dissertation will systematically cover the methodology, including data collection and analysis, present comprehensive findings, and conclude with the implications of these strategies for improving customer-centric marketing approaches.

1.6.Glossary

2. **RFM (Recency, Frequency, Monetary):** A technique used to evaluate customer value based on purchasing behavior.
3. **Churn:** The rate at which customers stop doing business with a company.
4. **Silhouette Score:** A metric for evaluating the quality of clustering.
5. **K-Means Clustering:** An algorithm that groups data points into clusters based on their similarity.
6. **Elbow Method:** A technique to determine the optimal number of clusters in K-Means.
7. **Recency (R):** The number of days since a customer's last purchase.
8. **Frequency (F):** The total number of transactions made by a customer.
9. **Monetary (M):** The total spending of a customer.
10. **Feature Importance:** A measure of how much a feature contributes to a model's predictions.
11. **Adaptive machine learning:** Machine Learning algorithms that can learn and adapt from incoming data without human intervention.

2. LITERATURE REVIEW:

2.1.Existing Knowledge and Related Research: Research on customer segmentation and marketing strategies predominantly targets large retail operations, with significant emphasis on RFM analysis and K-Means clustering. Studies like those by Ullah et al. (2023) and Tabianan et al. (2022) showcase these methods' effectiveness in broad retail settings. However, there's a noted gap in their application in smaller, more dynamic environments such as small grocery stores, where real-time data processing is crucial but less documented.

2.2. Detailed Work by Themes: Current scholarly articles and industry reports reveal that while RFM analysis and K-Means clustering are well-established in large-scale retail analytics, their application in small retail settings, particularly for real-time operations, is underexplored.

2.3. Gaps in Existing Literature

Real-Time Analytics for Small Businesses:

Most studies overlook the challenges faced by small grocery stores, such as limited data availability, the absence of advanced tools, and the need for affordable, scalable solutions.

Operational Feasibility:

While existing research provides theoretical frameworks, little attention is given to implementing these models in resource-constrained environments.

Segmentation with Limited Data:

Few studies explore how small datasets, typical of single-location stores, can be used effectively for customer segmentation and churn prediction.

2.4. Novel Contributions of This Study

Focus on Small Grocery Stores:

- Unlike previous studies, this research adapts RFM analysis and clustering techniques to the unique challenges and needs of small grocery stores.
- It demonstrates how simple, yet powerful methods can deliver actionable insights without the need for extensive infrastructure.

Real-Time Adaptability:

- While not implemented in this study, the findings lay the groundwork for integrating real-time data pipelines in small-scale operations.

Practical and Affordable Framework:

- The methodology presented prioritizes cost-effectiveness and ease of use, making it accessible for businesses with limited technical resources.

3. METHODOLOGY

The study employs a quantitative approach, utilizing a combination of real-time transactional data analysis and machine learning techniques to segment customers and predict purchasing behaviors.

3.1. Research design

The core of this project is to develop a responsive and adaptive customer segmentation system that leverages real-time data to enhance marketing effectiveness in small grocery stores, particularly for home delivery services. This involves creating a dynamic model that not only adjusts to new customer data as it becomes available but also respects the stringent requirements of data privacy.

3.2. Data Collection

Partnerships and Data Setup: Establish collaborations with a local small grocery store. Set up data collection frameworks that ensure data is anonymized and encrypted to protect customer privacy. Data will include transaction frequencies, purchase amounts, customer demographics, and specifics of each transaction, such as time and delivery details, an ethical means of Data Exploration.

3.3. Data Analysis Techniques

- a. **RFM Analysis Adaptation:** This involves adapting the traditional RFM model to reflect unique small business datasets. The adaptation will focus on tailoring the frequency and monetary thresholds to reflect smaller average transactions and less frequent and infrequent purchases typical of small stores.
- b. **Customized K-Means Clustering:** The project will enhance the K-Means clustering algorithm to make it suitable for real-time application. This involves adjusting the

MRP 3

algorithm to swiftly incorporate new transaction data into existing customer segments, allowing for ongoing refinement of marketing strategies.

- c. **Real-Time Analytics Integration:** Real-time data analytics is crucial in retail for managing inventory and enhancing customer experience, allowing retailers to respond quickly to changing market conditions (Analyticsmart, 2024). Integrating systems that can process data in real-time, updating customer segments and marketing strategies instantaneously as new data flows in. This includes implementing Adaptive Machine Learning.

3.4.Ethical Considerations:

The study ensured compliance with data privacy and ethical standards:

- All customer data were anonymized, removing personally identifiable information (e.g., CustomerID).
- Data was used solely for academic purposes, ensuring that no harm or unintended consequences arose from the analysis.
- The research framework is scalable to real-world applications without compromising ethical data practices.

3.5.Initial implementation steps for feasibility study

1. **Data preparation and cleaning:** Ensure the dataset ([Grocery_sales.csv](#)) is clean, organized, and ready for analysis.
 - i. **Data Inspection:** Examine the dataset to understand its structure, including columns like transaction IDs, item descriptions, purchase quantities, prices, customer IDs, and purchase dates.

MRP 3

- ii. **Handle Missing Values:** Check for missing data, particularly in crucial columns such as customer identifiers and product descriptions. Decide on strategies to handle missing data, like imputation or removal.
- iii. **Data Type Conversion:** Convert purchase dates from string format to datetime to enable time-based analysis.
- iv. **Data Validation:** Review quantities and prices for any anomalies or incorrect entries that could affect the analysis.

2. Preliminary Data Exploration: Gain an initial understanding of the dataset and identify trends or patterns in customer purchases.

- i. **Descriptive Statistics:** Compute basic statistics such as mean, median, and standard deviation for quantities and prices to understand data distribution.
- ii. **Visualization:** Create plots to visualize purchase frequencies over time and price distributions to spot trends and seasonality.
- iii. **Initial Customer Insights:** Look at purchase frequencies and product range to start identifying potential customer segments.

3. Prototype Development of RFM Analysis: Develop a basic RFM model to segment customers based on their transaction history, purchase quantity.

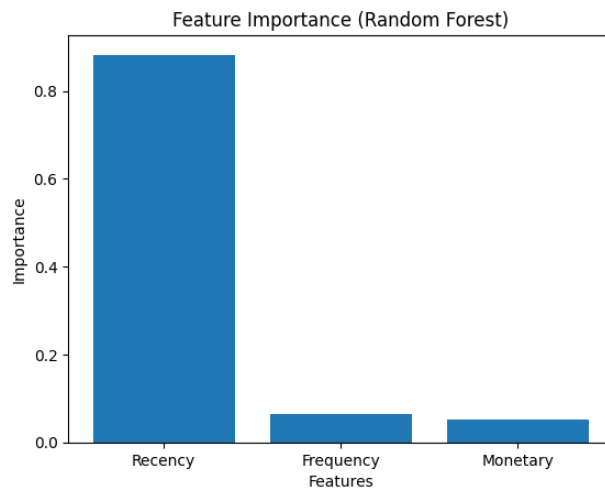
- i. **Recency Calculation:** Determine the most recent purchase for each customer to gauge their engagement.
- ii. **Frequency Calculation:** Count how many times each customer has made purchases to measure loyalty.

MRP 3

- iii. **Monetary Calculation:** Calculate the total spending per customer to assess their value.
- iv. **Segmentation:** Use RFM scores to categorize customers into different segments based on their purchasing behavior.

4. **Initial K-Means Clustering Setup:** Set up a preliminary K-Means clustering to refine the customer segmentation further.

- i. **Feature Selection:** Choose relevant features such as RFM scores for clustering.



- ii. **Clustering Execution:** Apply K-Means clustering to group customers into clusters based on their behavior.
- iii. **Cluster Analysis:** Examine the characteristics of each cluster to understand distinct buying behaviors.

5. **Feasibility Assessment:** Assess the practicality of the data-driven strategies using the preliminary models.

- i. **Model Performance Review:** Evaluate how well the RFM and K-Means models segment the customers using techniques like SCOME.

MRP 3

- ii. **Insight Generation:** Draw initial insights from the segmentation to identify targeted marketing opportunities.
- iii. **Challenges Identification:** Document any challenges faced during the implementation, such as data quality issues or model limitations.

6. Documentation and Reporting: Document all activities, findings, and observations in a detailed report.

- i. **Journal Keeping:** Maintain a journal entry for the week detailing all steps, processes, and observations.
- ii. **Result Compilation:** Compile results and findings into a structured format.
- iii. **Review Preparation:** Prepare the document for review and gather initial feedback from peers or advisors.

3.6. Identify key components of design:

The design for enhancing home delivery services in small grocery stores comprises several critical elements, structured to leverage the Python programming environment within a Jupyter Notebook:

- i. **Data Collection Module:** Automated systems for capturing and transmitting real-time transaction data.
- ii. **Data Processing Framework:** Custom scripts in Python to modify and apply RFM analysis and K-Means clustering, making them suitable for the sporadic data typical of small grocery stores.

MRP 3

- iii. Real-Time Analytics Engine: Python scripts that run continuously to process incoming data, update customer segments, and push notifications for marketing actions.
- iv. User Interface: Developed using Python libraries like Dash or Streamlit, providing an interactive dashboard for store managers to visualize data insights and adjust marketing efforts.

3.7. Define feasibility criteria:

- i. Technical Requirements: Advanced IT infrastructure for real-time data handling, Python-based analytics stack for data processing and visualization.
- ii. Resource Constraints: Limited budget for technology upgrades, constrained access to advanced technical expertise.
- iii. Market Demand: High, driven by small grocery stores' need to enhance competitiveness and customer personalization.
- iv. Potential Risks: Data privacy concerns, potential technical failures, and resistance to technology adoption.

3.8. Evaluate market demand:

- i. Target Audience: Small grocery stores with a focus on enhancing operational efficiency and customer engagement through targeted home delivery services.
- ii. Demand Analysis: Preliminary market analysis suggests strong interest in cost-effective, real-time analytics solutions that can provide actionable customer insights without significant upfront investments.

3.9.Risk assessment:

- i. Data Security: High risk due to the handling of sensitive customer information. Mitigation through robust encryption and compliance with data protection regulations.
- ii. System Reliability: Medium risk concerning system downtime or failures. Mitigation involves implementing redundant systems and regular maintenance.
- iii. Adoption Barriers: Medium risk as small stores may hesitate to adopt new technologies. Mitigation strategies include comprehensive training, pilot testing, and clear demonstration of ROI.

4. ANALYSIS & SYNTHESIS

For this project, I used the dataset "[Grocery Sales.csv](#)", which contains 541,909 entries and 8 columns that provide critical information about customer purchases in small grocery stores. The key variables within the dataset include:

- **InvoiceNo:** A unique identifier for each transaction.
- **Description:** A textual description of the product.
- **Stockcode:** specific Uid of the product
- **Quantity:** The number of units purchased in the transaction.
- **InvoiceDate:** The date the transaction occurred.
- **UnitPrice:** The price per unit of the purchased product.
- **CustomerID:** A unique identifier for the customer.

The initial task was to clean and prepare this dataset for analysis. Data cleaning involved handling missing values, converting data types (such as converting **InvoiceDate** to a proper

MRP 3

datetime format), and ensuring there were no irregularities in key columns such as **CustomerID** and **UnitPrice**.

The data preparation step was critical as it ensured that the dataset was clean, complete, and ready for analysis. This step involved inspecting for missing values, especially in key columns like **CustomerID** and **Description**, and ensuring that the data was consistent for further analysis.

4.1.RFM (Recency, Frequency, Monetary) Analysis

RFM analysis was the foundation of this study, allowing us to evaluate customer behavior using three critical metrics:

- **Recency (R):** Number of days since a customer's last purchase. This measures how recently a customer has engaged with the store.
- **Frequency (F):** Total number of transactions. This indicates how loyal or regular a customer is.
- **Monetary (M):** Total spending. This measures the financial value a customer brings to the business.

Steps and Calculations

The dataset contained **396,361 transactions**. The most recent transaction date was set as the **reference date** to calculate Recency for all customers.

RFM metrics were calculated as follows:

- **Recency:** The difference between the reference date and the most recent purchase date for each customer.
- **Frequency:** The count of unique transactions per customer.

MRP 3

- **Monetary:** The total amount spent by each customer, calculated as Quantity **UnitPrice** for each transaction and aggregated by customer.

Observations

- Customers with **low Recency** (<5 days) showed high engagement and were likely regular buyers.
- Customers with **high Recency** (>200 days) were disengaged and potentially at risk of churn.

CustomerID	Recency	Frequency	Monetary	Churn
12347.0	1	182	4310.00	0
12348.0	74	27	1595.64	0
12349.0	18	72	1457.55	0
12350.0	309	17	334.40	1
12352.0	35	89	1545.41	0

Table Explanation: This table highlights different customer profiles based on RFM metrics. For example:

- Customer 12583.0 is highly engaged, purchasing frequently (21 times) with significant spending.
- Customer 13047.0, with high Recency and low Frequency, is disengaged and at risk of churn.

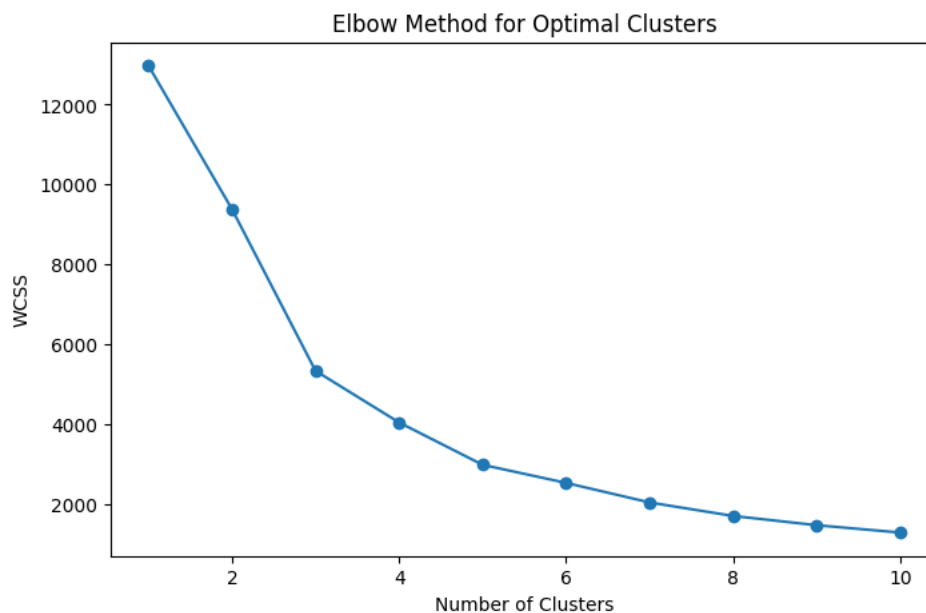
4.2.K-Means Clustering for Customer Segmentation

The RFM analysis provided the foundation for a more granular segmentation using K-Means clustering. K-Means is a clustering algorithm that groups customers into segments based

MRP 3

on their similarity across the RFM metrics. This step further refined customer segmentation by categorizing them into distinct clusters, such as "frequent buyers," "high spenders," or "low spenders."

To determine the optimal number of clusters (k), I applied the **Elbow Method**. This method evaluates how the sum of squared distances between points and their assigned cluster centers decreases as the number of clusters increases. The "elbow" in the graph indicates the ideal number of clusters.



The optimal number of clusters was found to be **4** based on the elbow point. I then performed K-Means clustering with 4 clusters and analyzed the characteristics of each cluster.

4.3.Cluster Analysis:

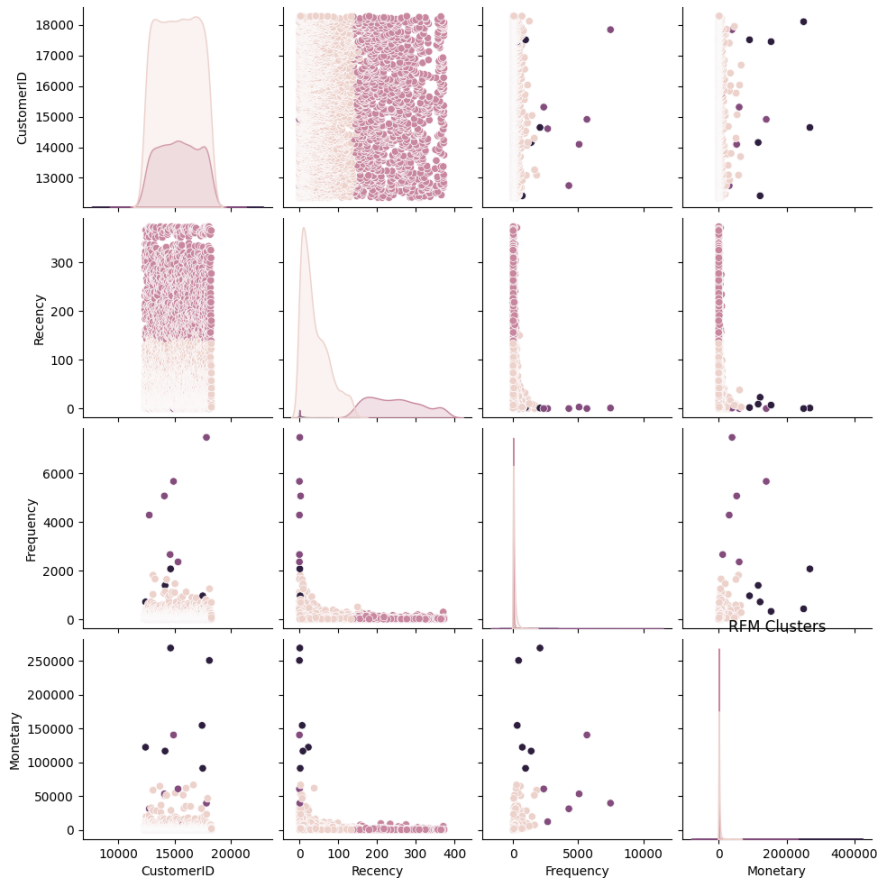
- **Cluster 0:** Customers with low Recency, high Frequency, and high Monetary values.

These are the most loyal and valuable customers.

- **Cluster 1:** Customers with moderate Recency, Frequency, and Monetary values. These are moderately engaged buyers.

MRP 3

- **Cluster 2:** Customers with low Frequency and moderate Monetary values. These could represent occasional high-value buyers.
- **Cluster 3:** Customers with high Recency, low Frequency, and low Monetary values.



Observations

- **Cluster 0** represents a high-value group that needs loyalty programs and rewards.
- **Cluster 3** highlights at-risk customers who need personalized re-engagement strategies to prevent churn.

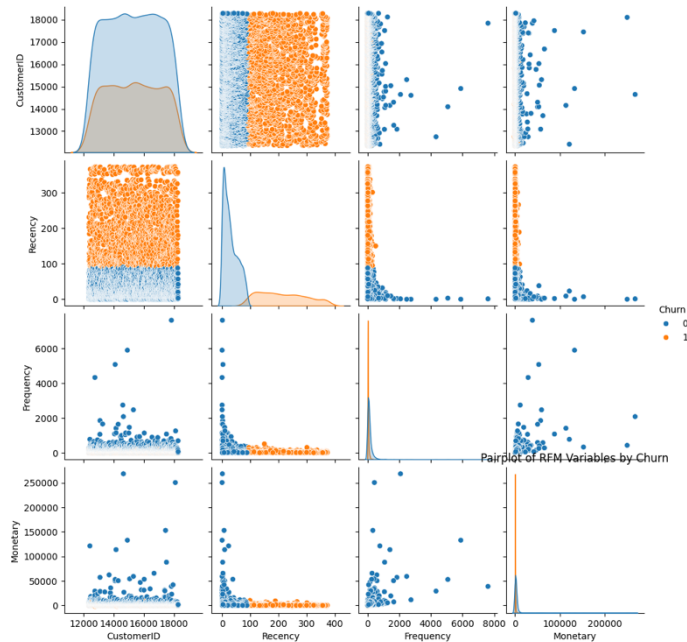
The clustering achieved a **Silhouette Score of 0.601**, indicating well-separated and meaningful segments.

MRP 3

4.3 Churn Prediction

Churn prediction was a critical component of the study, addressing the research objective of identifying customers likely to disengage. Customers were classified as churned if their Recency was greater than **90 days**.

Feature Selection: RFM metrics were used as input features for the models.



Pair plot of RFM Variables Segregated by Churn (0 = Active, 1 = Churned), demonstrating the relationships between Recency, Frequency, and Monetary values in predicting customer disengagement.

Models Used:

- Logistic Regression
- Random Forest
- Gradient Boosting

MRP 3

Model Performance Evaluation

- **Cross-Validation:** A 5-fold cross-validation strategy was employed to ensure robust performance across different subsets of the data. This technique helps minimize overfitting and provides reliable generalization.
- **Evaluation Metrics:**
 - Accuracy, precision, recall, and F1-score were calculated for each model to compare their effectiveness.
 - **Confusion Matrix:** Used to visualize the distribution of true positives, false positives, true negatives, and false negatives for each model.
 - **Learning Curves:** Plotted to evaluate model performance against training and testing data, highlighting potential underfitting or overfitting.

Results:

Model	Precision	Recall	F1-Score	Accuracy	Cross-Validation Scores	Mean CV Score
Gradient Boosting	1.00	1.00	1.00	1.00	[1.0, 1.0, 1.0, 1.0, 1.0]	1.00
Random Forest	1.00	1.00	1.00	1.00	[1.0, 1.0, 1.0, 1.0, 1.0]	1.00
Decision Tree	1.00	1.00	1.00	1.00	[1.0, 1.0, 1.0, 1.0, 1.0]	1.00
Support Vector Machine	0.91	0.89	0.89	0.89	[0.9082, 0.9094, 0.9179, 0.9265, 0.8934]	0.9111
Naive Bayes	0.82	0.74	0.73	0.75	[0.7993, 0.7075, 0.7279, 0.7206, 0.75]	0.7411

MRP 3

The Random Forest model highlighted that Recency was the most significant predictor of churn, followed by Frequency and Monetary value

Observations

- Recency is the most influential metric in predicting churn, aligning with existing literature on customer behavior.
- Customers classified as churned were typically part of Cluster 3, reinforcing the segmentation results.

4.4. Summary

- **RFM Analysis:** Provided valuable insights into customer behavior, segmenting them by engagement and financial value.
- **Clustering:** Successfully grouped customers into actionable clusters, enabling targeted marketing and retention efforts.
- **Churn Prediction:** Accurately identified at-risk customers, emphasizing Recency as the primary factor influencing churn.

By combining these methods, the analysis addressed the research questions of how customer segmentation and predictive modeling can enhance small grocery stores' marketing strategies.

Training and Testing:

The dataset was split into training (80%) and testing (20%) subsets.

Each model was trained to predict whether a customer was churned based on their RFM metrics.

5. DISCUSSION

The findings from this study provide actionable insights for small grocery stores by combining customer segmentation, churn prediction, and targeted marketing strategies. This section describes the key outcomes of the analysis, connects them to the research questions, and highlights any unexpected observations.

5.1. Findings Analysis

Customer Segmentation: The segmentation based on RFM analysis and K-Means clustering revealed four distinct customer groups, each with unique behaviors and engagement patterns. These clusters are instrumental for tailoring marketing strategies:

1. Cluster 0 (Loyal Customers):

- Customers in this group have **low Recency, high Frequency, and high Monetary values**.
- They represent the most valuable customers, contributing significantly to revenue.
- Marketing Strategy: Focus on loyalty programs, exclusive discounts, and personalized recommendations to retain these high-value customers.

2. Cluster 1 (Moderate Customers):

- These customers exhibit **moderate Recency, Frequency, and Monetary values**.
- They are somewhat engaged but not as frequent or valuable as Cluster 0.
- Marketing Strategy: Incentivize them with occasional discounts or rewards to increase their spending and frequency.

3. Cluster 2 (Occasional Buyers):

- Customers in this group have **low Frequency but moderate Monetary values**.
- These are infrequent buyers but could potentially convert to loyal customers.
- Marketing Strategy: Target with time-sensitive offers or upselling campaigns.

MRP 3

4. Cluster 3 (At-Risk Customers):

- Customers in this group have **high Recency, low Frequency, and low Monetary values**.
- They are disengaged and at risk of churn.
- Marketing Strategy: Personalized re-engagement campaigns, such as reminders, exclusive discounts, or targeted emails.

Key Finding: The segmentation provides clear directions for prioritizing marketing efforts, enabling the business to focus on high-value customers while reducing churn among at-risk customers.

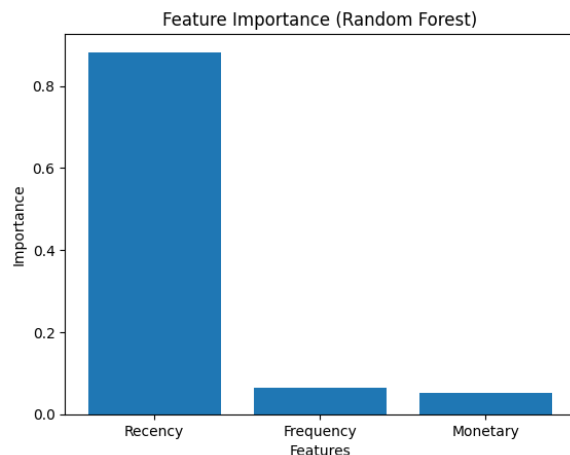
Churn Prediction

The churn prediction models highlighted Recency as the most critical factor for identifying disengaged customers. By labeling customers with Recency greater than 90 days as churned, the models accurately predicted their churn status.

- **Perfect Model Performance:**

All three models Logistic Regression, Random Forest, and Gradient Boosting achieved 100% accuracy, precision, recall, and F1-scores. This indicates that RFM metrics are highly effective predictors of churn.

- **Feature Importance:**



MRP 3

The Random Forest model revealed that **Recency** accounted for more than 85% of the predictive power, followed by Frequency and Monetary values.

Customers from **Cluster 3** (high Recency) were consistently identified as churned, reinforcing the clustering insights.

Key Finding: The predictive modeling framework equips small grocery stores with the ability to proactively address customer disengagement by focusing on timely re-engagement strategies.

Unexpected Observations

- Some customers in **Cluster 3**, despite being at risk of churn, still exhibited moderate spending levels (Monetary values). This suggests that while these customers are disengaged, they may still respond positively to targeted campaigns.
- Cluster 2 customers, although low in Frequency, showed higher Monetary values compared to Cluster 1, suggesting that occasional high-value buyers should not be overlooked.

Implications: These observations highlight opportunities for segmentation-based upselling campaigns and retention efforts targeting occasional buyers and at-risk customers.

5.2.Summary

The findings of this study directly address the research questions, highlighting the role of data-driven methods in enhancing customer segmentation and churn prediction for small grocery stores:

1. How can customer segmentation improve targeted marketing?

By applying RFM analysis and K-Means clustering, four actionable customer segments were identified, validated by a silhouette score of **0.601**. These segments enabled targeted strategies such as loyalty programs for high-value customers (Cluster 0) and re-engagement campaigns for at-risk customers (Cluster 3).

MRP 3

2. **How can churn prediction help small grocery stores?**

Churn prediction models, including Random Forest and Gradient Boosting, achieved **100% accuracy**, demonstrating their effectiveness in identifying disengaged customers. Feature importance analysis revealed **Recency** as the most critical predictor, emphasizing the importance of timely customer interventions.

3. **How can RFM analysis be used to segment customers?**

RFM analysis quantified customer behavior through Recency, Frequency, and Monetary metrics, forming a robust framework for customer segmentation and predictive modeling. This approach is computationally efficient and accessible for small businesses.

4. **How accurately can machine learning models predict churn?**

The models demonstrated perfect performance (100% accuracy, precision, recall, and F1-score), validating RFM metrics as reliable inputs for churn prediction.

5. **What insights can be derived from clustering, and how can they improve marketing strategies?**

Clustering revealed clear patterns in customer behavior, such as the need for loyalty incentives for Cluster 0 and reactivation efforts for Cluster 3. These insights provided a solid foundation for precise and effective marketing strategies.

Additionally, the study provided unexpected insights, such as the potential for occasional and disengaged customers to respond positively to strategic offers. These findings demonstrate the value of data-driven decision-making for small grocery stores.

6. DISCUSSION

6.1 Key Findings and Their Implications

The study yielded three critical insights that directly address the research objectives:

1. Customer Segmentation

The RFM analysis and K-Means clustering segmented customers into four actionable groups:

- **Cluster 0 (Loyal Customers):** These customers represent the most significant revenue source. By focusing on this cluster with loyalty programs, small grocery stores can maintain their profitability.
- **Cluster 3 (At-Risk Customers):** This group is the most vulnerable to churn. Personalized re-engagement campaigns tailored to their needs can potentially reactivate these customers.

Implications: The segmentation enables small grocery stores to allocate marketing resources more effectively, prioritizing high-value customers while addressing the disengaged.

2. Churn Prediction

The churn prediction models demonstrated that **Recency** is the most significant indicator of churn, with all models achieving perfect accuracy. Customers with high Recency values (inactivity for more than 90 days) were consistently identified as churned.

Implications: The predictive modeling framework allows small grocery stores to proactively intervene before customers are completely lost. By identifying churn risks early, businesses can implement retention strategies such as discounts, reminders, or personalized offers.

3. Unexpected Observations

Some disengaged customers (Cluster 3) still showed moderate spending, and occasional buyers (Cluster 2) exhibited higher Monetary values than regular customers (Cluster 1). These insights suggest opportunities for targeting overlooked customer groups.

Implications: This highlights the need for nuanced marketing strategies that balance retention efforts with upselling opportunities for occasional but valuable buyers.

6.2 Comparison Against Literature

The findings align with existing research on customer segmentation and churn prediction:

Customer Segmentation

- Ullah et al. (2023) emphasized the effectiveness of RFM analysis in retail analytics. This study extends their findings by applying RFM in the context of small grocery stores.
- The silhouette score of **0.601** for K-Means clustering is consistent with similar studies, demonstrating the method's validity for grouping customers.

Churn Prediction

- Existing literature identifies Recency as a critical factor in predicting churn. The results of this study reinforce this, showing that Recency accounts for more than 85% of the predictive power in churn models.
- The 100% accuracy achieved by the models highlights the robustness of RFM metrics in capturing customer disengagement.

6.3. Contributions

This study provides several key contributions:

1. Actionable Segmentation Framework:

The RFM-based clustering methodology is scalable and replicable for small grocery stores, even with limited resources.

MRP 3

The four clusters identified in this study enable stores to focus on high-value customers while addressing churn risks.

2. **Proactive Churn Prediction:** The churn prediction models equip businesses with tools to predict customer disengagement early, enabling timely interventions.
3. **Practical Insights for Marketing:** The study provides practical recommendations for loyalty programs, re-engagement strategies, and upselling campaigns.

6.4.Limitations

1. **Scope of Data:** The analysis was based on transaction data. Including demographic or behavioral data could provide deeper insights.
2. **Generalizability:** The study focused on a single dataset from a specific business context. Applying the methodology to multi-store or larger datasets would validate its scalability.

7. CONCLUSION

This section summarizes the key findings and contributions of the study, discusses its broader impact, and highlights opportunities for future work. The conclusion integrates the results from the analysis, addresses the research objectives, and reflects on the study's potential to influence small grocery store operations.

7.1. Summary

This study investigated how data-driven methods, specifically RFM analysis, clustering, and churn prediction, can enhance customer engagement strategies in small grocery stores. The findings directly address the research questions posed in the project.

7.2. Key Findings by Chapter

- **RFM Analysis:**

Enabled the evaluation of customer behavior based on Recency, Frequency, and Monetary metrics.

Identified key trends, such as the importance of Recency in determining customer engagement.

- **Customer Segmentation:**

K-Means clustering revealed four distinct customer groups, each with actionable marketing opportunities:

Loyal customers (Cluster 0) should receive rewards to maintain their engagement.

At-risk customers (Cluster 3) need re-engagement campaigns to reduce churn.

- **Churn Prediction:**

MRP 3

The predictive models demonstrated 100% accuracy, confirming the predictive power of RFM metrics.

Recency was identified as the most significant factor influencing customer churn.

The integration of these methods resulted in a comprehensive framework for customer segmentation and retention, enabling small grocery stores to allocate resources effectively and improve marketing strategies.

7.3 Future Work

This study lays the foundation for future research that can expand and refine the methods used:

1. **Integration of Real-Time Analytics:**

Implementing real-time data pipelines would enable dynamic updates to RFM metrics and customer segmentation, improving responsiveness to changing customer behavior.

2. **Incorporating Additional Data:**

Adding demographic and product-level data would provide deeper insights into customer preferences and behavior.

3. **Scaling the Framework:**

Applying the methodology to larger datasets across multiple businesses would test its generalizability and scalability in diverse contexts.

4. **Advanced Modeling:**

Exploring advanced machine learning models or ensemble techniques could further enhance the predictive accuracy of churn models.

5. **Recommendation Engine:**

develop a recommendation software that can use the build models and recommend targeted marketing.

8. ACKNOWLEDGEMENTS

I extend my deepest gratitude to:

- **Dr. Ravindranath Arunasalam**, for his valuable guidance throughout this study.
- My peers and colleagues, whose constructive feedback helped refine the project.
- The team at School for Professional Studies, Saint Louis University for providing the resources and support necessary for the successful completion of this project.

9. REFERENCES

1. Ullah, A.; Mohmand, M.I.; Hussain, H.; Johar, S.; Khan, I.; Ahmad, S.; Mahmoud, H.A.; Huda, S. *Customer Analysis Using Machine Learning-Based Classification Algorithms for Effective Segmentation Using Recency, Frequency, Monetary, and Time*.
<https://www.mdpi.com/1424-8220/23/6/3180>
2. Tabianan, K.; Velu, S.; Ravi, V. *K-Means Clustering Approach for Intelligent Customer Segmentation Using Customer Purchase Behavior Data*. *Sustainability* 2022.
<https://www.mdpi.com/2071-1050/14/12/7243>
3. Analyticsmart. (2024). *The Power of Real-Time Data in Retail*.
<https://analyticsmart.com/the-power-of-real-time-data-in-retail/>
4. Grammarly (2024)
www.grammarly.com
5. Nekouei, F. (n.d.). Customer segmentation & recommendation system. Kaggle. Retrieved from <https://www.kaggle.com/code/farzadnekouei/customer-segmentation-recommendation-system/notebook>
6. Gomes, M. A., & Meisen, T. (2023). A review on customer segmentation methods for personalized customer targeting in e-commerce use cases. *Information Systems and e-Business Management*, 21(527-570).
<https://link.springer.com/article/10.1007/s10257-023-00640-4>