Customer Retention Analysis and Behavioral Insights for a Mid-Sized Retail Store

1. Business Problem:

The retail store was experiencing a decline in customer retention, which impacted both profitability and market share despite ongoing loyalty programs and promotions. Addressing this issue required a deep understanding of why customers churned and the development of strategies to retain at-risk segments. This project aimed to thoroughly analyze customer purchasing behavior, highlight segments likely to disengage, and develop targeted retention strategies. By identifying key churn indicators, such as low purchase frequency and specific product preferences, the store could employ data-driven initiatives to enhance customer loyalty, maintain valuable relationships, and support sustainable business growth.

2. Background:

Customer retention is essential for retail success, as retaining existing customers is often more cost-effective than acquiring new ones. In mid-sized retail stores, customer loyalty directly impacts revenue growth, yet many businesses face difficulties in maintaining strong retention rates. While loyalty programs and promotions are frequently used, businesses often lack a clear understanding of the factors driving customer churn. Elements such as purchase frequency, product satisfaction, and income levels significantly influence retention, but traditional analysis methods may overlook nuanced patterns in customer behavior, limiting the potential for effective retention improvements.

This project utilizes advanced analytics and machine learning to examine customer retention in greater detail. Using a robust dataset from Kaggle with over 302,000 records on customer demographics, transactions, product types, and satisfaction, the analysis assesses churn risk through techniques like K-means clustering, association rule mining, and churn prediction models. By segmenting customers based on their behavior, this project identifies groups most likely to churn, focusing on characteristics such as low purchase frequency or specific product preferences. Additionally, factors influencing churn—such as reduced purchasing activity or lower product ratings—are explored to provide actionable insights.

These insights guide targeted strategies to boost customer loyalty, such as personalized offers for low-engagement customers, rewards for high-spending clients, and tailored promotions for at-risk segments. By proactively addressing high-risk behaviors and preferences, the retail store can enhance customer retention, fostering long-term loyalty and mitigating the impact of churn on business performance.

3. Data Explanation:

The dataset, sourced from Kaggle, consists of 302,011 rows and 30 columns, providing a comprehensive view of customer transactions, demographics, and product-related details. Key variables include demographics such as age, gender, and income; transaction details like purchase dates, total purchases, and customer feedback; as well as product attributes covering category, brand, and type. This rich dataset allows for an in-depth analysis of customer behavior, helping to identify patterns in purchasing, assess churn risk, and gauge engagement levels.

The dataset's extensive range of variables enables a robust exploration of factors influencing customer loyalty and churn. By leveraging this information, the analysis builds a solid foundation for developing tailored retention strategies and targeted marketing interventions aimed at improving customer satisfaction and reducing churn rates.

4. Methods:

The analysis employed a structured approach to understand customer behavior and identify churn risk. Data preprocessing involved cleaning, handling missing values, and adding features like average order frequency for a more insightful analysis. In the Exploratory Data Analysis (EDA) phase, descriptive statistics and visualizations outlined trends in purchasing, demographics, and feedback, alongside correlation analysis to identify factors influencing churn.

Three core modeling techniques were applied: K-means clustering to segment customers by purchase behavior and demographics; decision tree classifiers for predicting churn risk; and association rule mining to find frequent patterns in transactions and feedback. Clustering was evaluated using silhouette scores, while churn prediction models were validated on accuracy, precision, and recall metrics. Association rules were assessed by support, confidence, and lift. This combined approach generated actionable insights, guiding targeted strategies to boost customer retention and engagement.

5. Results:

The analysis identified key patterns and factors impacting customer retention, offering actionable insights for boosting engagement.

Question 1: Monthly purchasing trends showed clear seasonality. March and July were high-activity months, with purchases in March increasing from 1,686 in 2022 to 1,910 in 2023, and July showing a significant rise from 1,560 to 2,003. Conversely, February and August experienced declines, underscoring the need for seasonally adaptive strategies (Fig.1).

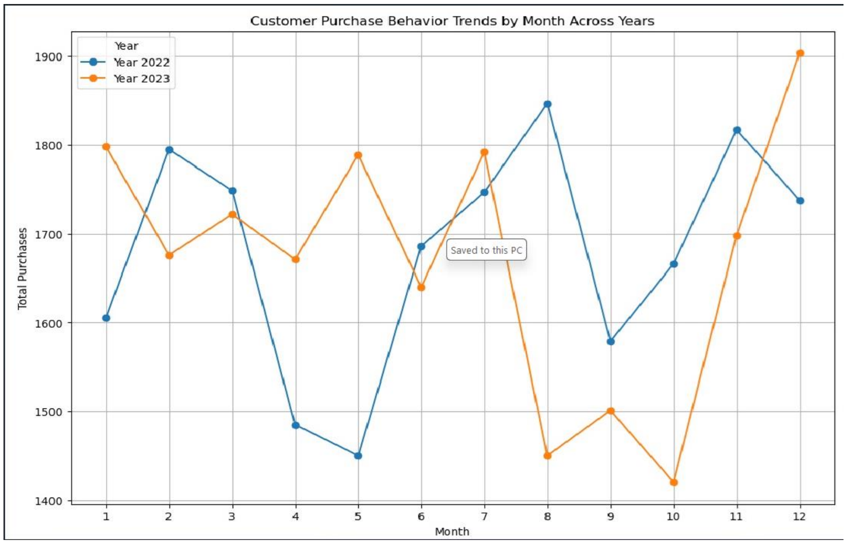

Fig. 1. Customer purchase behavior trend

Question 2: Purchase frequency was the strongest churn predictor, with at-risk customers averaging 2.51 monthly purchases, compared to 8.37 for engaged customers. This discrepancy highlights a clear link between low engagement and churn risk, suggesting that encouraging higher purchase frequency could mitigate churn (Fig.2).

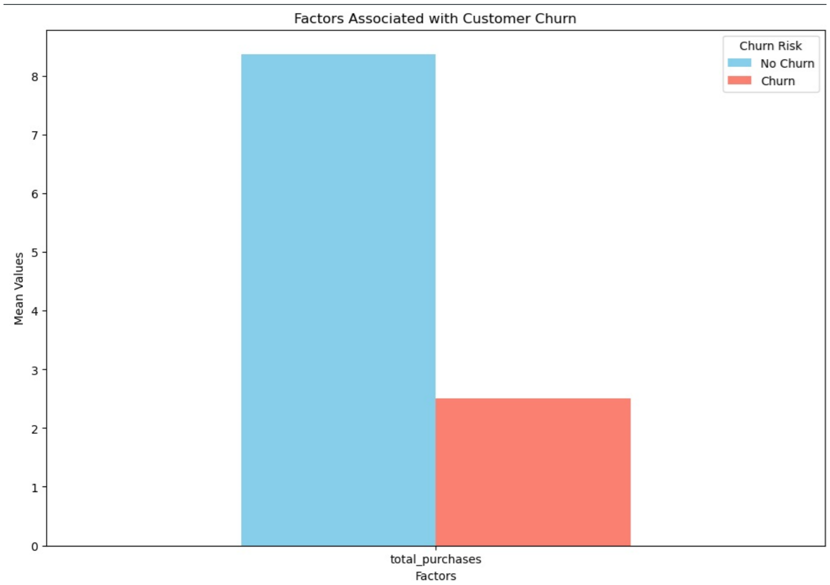

Fig. 2. Factors associated with customer churn

Question 3: Customer segmentation identified Segment 1 as the most vulnerable to churn, with a 1.0 risk and only 2.46 average monthly purchases. Segment 2 had a moderate churn risk of 0.34, averaging 5.34 purchases, while Segment 0 showed no churn risk, with an average of 6.80 purchases (Fig.3). These insights allow for precise targeting of retention strategies.

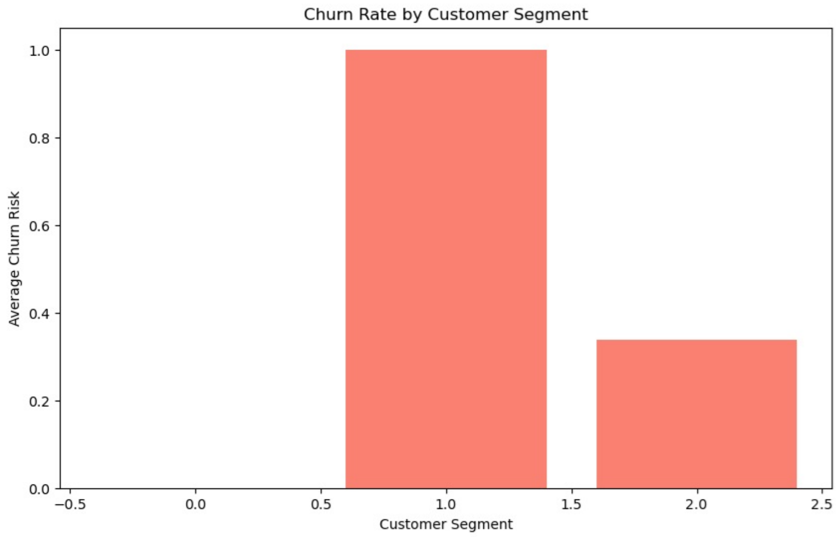

Fig.3. Churn rate by customer segment

Question 4: For high-risk customers, Electronics (23,642 purchases) and Grocery (22,104) were top categories, revealing opportunities for targeted promotions. Focusing on these preferences enables the development of specialized marketing efforts aimed at re-engaging at-risk customers (Fig.6).

In sum, the results underscore the importance of segmentation, seasonality, and product preferences for effective engagement strategies, providing a clear path for improving customer retention.

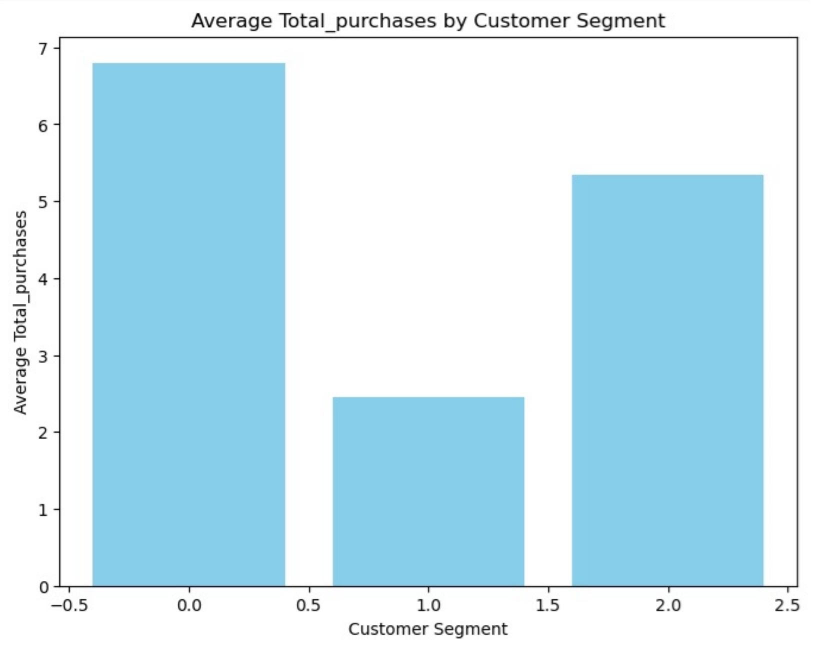

Fig. 4. Average total purchases by customer segment

6. Discussion of Analysis Results:

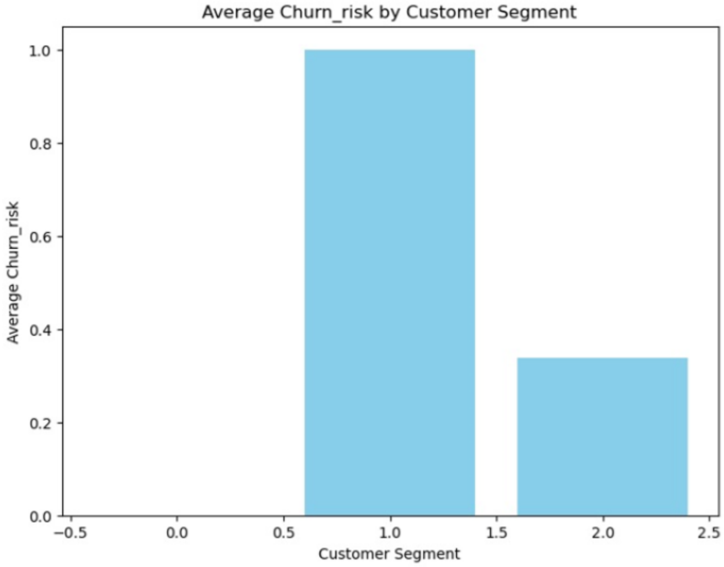

Fig. 5. Average churn risk by customer segment

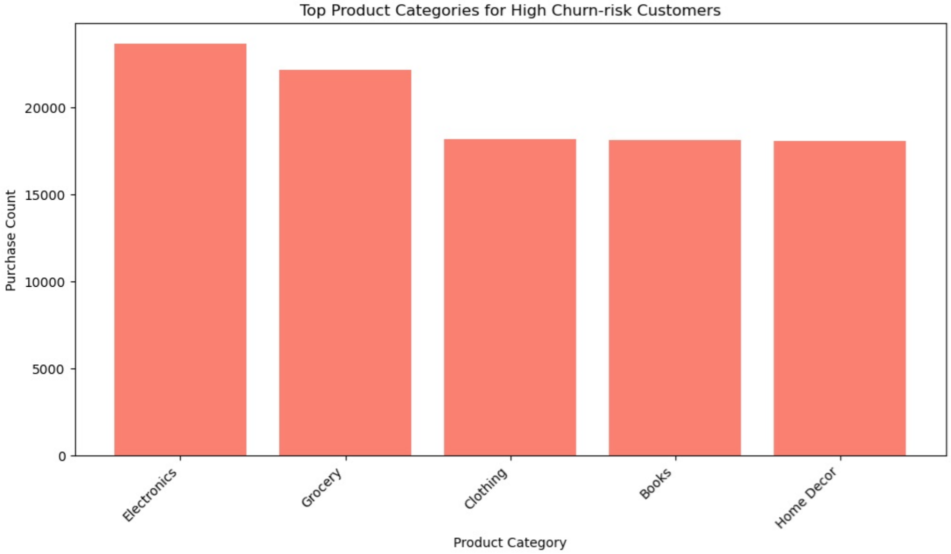

Fig. 6. Top product categories for high churn risk customers.

The analysis of customer behavior, churn factors, and segmentation yielded valuable insights that support actionable strategies for boosting customer retention and engagement.

Purchasing Behavior Trends:

The line plot of monthly purchases over 2022 and 2023 reveals clear seasonal trends. Notable peaks were observed in March and July, with purchases increasing from 1,686 to 1,910 and from 1,560 to 2,003, respectively, between these years. These months represent opportunities for enhanced marketing and inventory planning to meet higher demand. Conversely, declines in February and August—1,657 to 1,532 and 1,606 to 1,483—suggest potential for seasonal promotions to drive activity in slower periods. Recognizing and adapting to these patterns enables the retail store to align staffing, stock levels, and promotional campaigns with seasonal customer behaviors.

Key Factors in Customer Churn:

Churn analysis identified low purchase frequency as a strong churn indicator. Customers flagged as "at-risk" averaged only 2.51 monthly purchases, compared to 8.37 for engaged customers. This stark difference underscores the correlation between low engagement and churn risk. By establishing this frequency threshold, the store can create retention strategies aimed at re-engaging at-risk customers. Personalized promotions, loyalty rewards, or exclusive offers could increase purchase frequency among low-engagement customers, ultimately helping to lower churn and increase customer loyalty.

Customer Segmentation and Churn Patterns:

Using KMeans clustering, the analysis segmented customers into three groups with distinct churn and purchasing characteristics. Segment 1 has the highest churn risk (1.0), averaging 2.46 monthly purchases, making it the primary focus for re-engagement. Segment 2 exhibits a moderate churn risk (0.34) with an average of 5.34 purchases, indicating improvement potential. Segment 0, with no churn risk and an average of 6.80 purchases, represents a loyal customer base. Tailored marketing efforts, such as offering incentives to high-churn segments and maintaining engagement with moderately at-risk segments, are recommended based on these findings.

Product Preferences for High Churn-Risk Customers:

High churn-risk customers show strong preferences for categories like Electronics (23,642 purchases) and Grocery (22,104), suggesting these are key areas for targeted retention strategies. Popular categories also include Clothing (18,126), Books (18,106), and Home Decor (18,028), which offer opportunities for specialized marketing, such as bundled deals or discounts. Targeting these categories allows the store to better engage at-risk customers by focusing on their product preferences, ultimately supporting retention and improved customer satisfaction.

7. Conclusions:

The analysis provided key insights into improving customer retention by examining purchasing behaviors, churn indicators, and customer segmentation. Low purchase frequency, income levels, and specific product interests were found to significantly correlate with churn risk. Customers who purchase more frequently demonstrated stronger loyalty, emphasizing the effectiveness of targeted rewards to boost their engagement. Segmentation analysis revealed high-risk groups, particularly among low-frequency shoppers, who could benefit from tailored retention efforts. By implementing focused strategies—such as personalized discounts, exclusive loyalty programs, and seasonal promotions aligned with customer preferences—the retail store can effectively reduce churn, improve customer satisfaction, and foster long-term loyalty for sustainable growth.

8. Assumptions:

The analysis was based on several key assumptions. It assumed that the dataset accurately reflects the store's overall customer demographics and purchasing behaviors. It also presumed that the identified relationships between customer feedback, purchase frequency, and churn risk remain consistent over time. Additionally, the machine learning models used, such as K-means clustering and decision trees, were assumed to effectively capture relevant patterns to provide reliable segmentation and retention insights. Lastly, it was assumed that external influences, like market fluctuations or economic changes, had minimal impact on the analysis results.

9. Limitations:

This project faced several limitations, primarily due to its dependence on historical transaction data, which may not capture recent shifts in customer behavior influenced by external factors like competitor strategies or market fluctuations. Additionally, data inconsistencies and missing values could affect the accuracy of the analysis and model predictions. While K-means clustering and decision trees were effective for segmentation and churn analysis, they might overlook nuanced patterns better detected by more sophisticated algorithms. Furthermore, the analysis centered on economic and transactional data, excluding behavioral or qualitative factors that could provide a deeper understanding of customer retention and engagement dynamics.

10. Challenges:

The project encountered several challenges, notably in managing large datasets with missing or inconsistent data, which affected data quality and analysis accuracy. Defining customer churn was complex, as selecting the right threshold influenced the outcomes significantly. Utilizing traditional models like K-means clustering and decision trees restricted the depth of pattern detection, possibly missing more nuanced trends. Additionally, the reliance on historical data posed limitations, as it might not fully capture recent shifts in customer behavior influenced by market dynamics or competitive actions.

11. Future Uses/Additional Applications:

This project provides a foundation for expanding into predictive modeling and real-time analytics to better understand customer behavior. Leveraging more advanced machine learning algorithms could improve churn prediction accuracy and reveal deeper insights into customer lifetime value. Integrating external data sources, such as competitor pricing or social media sentiment, could offer a comprehensive view of customer loyalty and market trends. Additionally, this analytical framework can be adapted to other industries, including hospitality and e-commerce, optimizing customer retention strategies and driving engagement across diverse sectors.

12. Recommendations:

To enhance customer retention, the following targeted strategies are suggested:

Engage At-Risk Customers: Offer personalized discounts, promotions, or loyalty rewards to low-income customers and those with low purchase frequency to boost engagement and increase their purchasing activity.

Monitor Customer Feedback: Actively analyze and respond to customer feedback, focusing on at-risk customers, to address dissatisfaction early and improve the shopping experience.

Loyalty Rewards for High-Spending Customers: Establish exclusive loyalty programs for high-income and high-frequency customers to strengthen brand loyalty and incentivize further purchases.

Targeted Promotions for High Churn-Risk Categories: Leverage insights from high churn-risk categories, like Electronics and Grocery, to provide relevant promotions and rekindle interest among disengaging customers.

13. Implementation Plan:

The implementation plan includes a phased approach:

Week 9: Final Preparations and Model Customization

Data Finalization: Clean and prepare the final dataset based on analysis requirements.
Model Customization: Fine-tune churn prediction and customer segmentation models for retail-specific needs.

Stakeholder Training: Conduct brief sessions for stakeholders on model usage, data insights, and decision-making.

Week 10: Pilot Testing and Integration

Pilot Deployment: Deploy the churn prediction model on a small customer sample for testing.
Feedback Loop: Collect feedback from team members, noting model performance and operational impact.
Refinement: Make initial adjustments to models and processes based on pilot feedback.

Week 11: Full-Scale Model Deployment

Implementation Across All Data: Deploy churn and segmentation models across the entire dataset.
Monitor Performance: Track model accuracy, customer response, and retention rates in real time.
Troubleshooting: Identify and resolve integration issues promptly.

Week 12: Evaluation and Optimization

Review KPIs: Analyze key performance indicators such as churn rate reduction and customer engagement levels.
Adjust Strategies: Refine retention strategies based on early outcomes and team insights.
Continuous Monitoring Setup: Set up automated tracking and reporting systems for ongoing analysis.

14. Ethical Assessment:

To ensure ethical and responsible data usage, I focused on the following principles:

Data Privacy and Security: I ensured sensitive customer data, such as demographics and purchasing behavior, was protected with robust storage and access protocols to prevent unauthorized access and misuse.

Avoiding Analytical Bias: I conducted regular reviews of model assumptions to avoid biases in customer segmentation and retention strategies, ensuring fair treatment of all customer groups.

Transparency in Data Use: I prioritized transparency by clearly communicating how customer data would be utilized, particularly in personalized marketing, to foster trust and avoid misuse.

Customer-Centric Decision-Making: I developed strategies aimed at genuinely enhancing customer satisfaction, avoiding practices that solely prioritize profit over customer welfare.

15. References

Hsu, H. Y., & Lee, C. L. (2021). Predictive analytics in smart agriculture. Routledge. https://doi.org/10.4324/9780367424218

Chen, H., Chiang, R. H., & Storey, V. C. (2012). Business intelligence and analytics: From big data to big impact. MIS Quarterly, 36(4), 1165-1188. https://doi.org/10.2307/41703503

Kumar, V., & Shah, D. (2004). Building and sustaining profitable customer loyalty for the 21st century. Journal of Retailing, 80(4), 317-330. https://doi.org/10.1016/j.jretai.2004.10.007

Kaggle. (n.d.). Retail analysis large dataset. Retrieved from https://www.kaggle.com/dsv/8693643

Lejeune, M. A. (2001). Measuring the impact of data mining on churn management. Internet Research, 11(5), 375-387. https://doi.org/10.1108/10662240110410198

Neslin, S. A., Gupta, S., Kamakura, W., Lu, J., & Mason, C. H. (2006). Defection detection: Measuring and understanding the predictive accuracy of customer churn models. Journal of Marketing Research, 43(2), 204-211. https://doi.org/10.1509/jmkr.43.2.204

Ngai, E. W. T., Hu, Y., Wong, Y. H., Chen, Y., & Sun, X. (2011). The application of data mining techniques in financial fraud detection: A classification framework and an academic review of literature. Decision Support Systems, 50(3), 559-569. https://doi.org/10.1016/j.dss.2010.08.006

Rosset, S., Neumann, E., Eick, U., & Vatnik, N. (2003). Customer lifetime value models for decision support. Data Mining and Knowledge Discovery, 7(3), 321-339. https://doi.org/10.1023/A:1024950200689