INFO BHARAT INTERNS

(Internship Program)

REPORT ON

CUSTOMER AND SALES DATA SET

Under the Guidance of:

info bharat interns

Date of Submission:

13 july, 2025

Submitted by:

Saurabh Kumar

B.Tech (CSE)

shivabtech2022@gmail.com

Table of Contents

Executive Summary

This comprehensive report presents an in-depth multi-faceted analysis of customer and sales data, aimed at uncovering critical insights that can drive strategic decision-making, boost revenue growth, and enhance customer retention. Leveraging advanced data analytics techniques—including robust data preprocessing, feature engineering, customer segmentation, time-series forecasting, churn prediction, and market basket analysis—this study offers a holistic view of customer behavior and sales dynamics.

Through meticulous data cleaning and enrichment, key metrics such as Customer Lifetime Value (CLV), Recency, Frequency, and Monetary scores were derived to form the backbone of deeper analyses. Exploratory Data Analysis (EDA) revealed significant patterns, including seasonal fluctuations in sales, demographic trends influencing purchase behavior, and correlations between loyalty scores and overall customer value.

Advanced clustering algorithms like Gaussian Mixture Models and Agglomerative Clustering enabled the identification of distinct customer segments, ranging from high-value loyalists to dormant customers. Each segment was profiled with tailored strategies, ensuring marketing efforts can be precisely targeted for maximum impact.

Utilizing the Prophet model, future sales trends were forecasted, highlighting periods of expected growth and potential dips. These forecasts are invaluable for aligning marketing campaigns, inventory planning, and resource allocation. Meanwhile, churn prediction was executed using an optimized XGBoost classifier, pinpointing customers at high risk of attrition and providing actionable opportunities to intervene through personalized retention efforts.

Market basket analysis further uncovered product affinities and frequent co-purchases, opening the door for effective cross-selling and bundling strategies. Such insights are crucial for increasing average order values and enhancing the customer shopping experience.

In conclusion, this analysis serves as a strategic blueprint for data-driven decision-making. By focusing on high-value customer segments, proactively mitigating churn, leveraging product associations for cross-selling, and aligning operational efforts with forecasted demand, the business is positioned to achieve sustained growth and build deeper, more profitable customer relationships.

Data Preprocessing & Feature Engineering

Effective data-driven analysis begins with rigorous preprocessing and thoughtful feature engineering. In this project, multiple advanced techniques were employed to ensure the dataset was both clean and analytically rich, enabling robust insights and high-performing predictive models.

Handling Missing Values

The raw dataset contained missing entries across critical variables such as Loyalty Score, Age, and transaction-level details. To address this, an **Iterative Imputer (Multiple Imputation by Chained Equations - MICE)** was used, which intelligently predicts missing values by modeling each feature as a function of others. This approach preserves underlying data relationships far better than simple mean or median imputation. Additionally, columns with near-zero variance or excessive nulls were carefully evaluated and dropped if necessary to avoid distorting subsequent analyses.

Outlier Detection & Treatment

Outliers can significantly skew model performance and obscure genuine patterns. The **Interquartile Range (IQR) method (Tukey's rule)** was applied to detect extreme values across numerical features. Instead of deleting data, which risks losing valuable information, outliers were capped at calculated upper and lower bounds. This ensures the integrity of the data distribution while minimizing undue influence on clustering and forecasting.

Data Normalization & Scaling

To prepare data for algorithms sensitive to scale (such as K-means clustering, Gaussian Mixture Models, and distance-based methods), **Min-Max scaling** was applied, transforming features to a common 0–1 range. For models requiring standardized inputs, such as regression and tree-based techniques, **StandardScaler** transformations were also tested to ensure model robustness.

Feature Engineering

Beyond cleaning, new variables were engineered to deepen insights and power advanced analytics:

- **Customer Lifetime Value (CLV):** Calculated as the aggregate monetary spend over the observed period, serving as a critical metric for customer profitability.
- **Recency:** Measured as days since the customer's last purchase, a key indicator of engagement.

- **Frequency:** Total number of purchases made by each customer, highlighting behavioral intensity.
- **Recency & Frequency Scores:** Created using quantile-based ranking (0–3 scale) to simplify segmentation.
- **Average Purchase Frequency:** Derived to capture typical transaction intervals.
- **Discount Utilization Rate:** Percent of purchases where discounts were applied, helping profile price sensitivity.

These engineered features formed the backbone of the customer segmentation models and predictive churn analysis. They also served to uncover hidden patterns in buying behavior and loyalty that raw transactional data alone could not reveal.

Final Prepared Dataset

The resulting dataset after preprocessing and feature engineering was well-structured, free of major anomalies, and enriched with behavioral metrics—ready to power sophisticated segmentation, forecasting, and predictive modeling.

Exploratory Data Analysis (EDA)

A comprehensive Exploratory Data Analysis (EDA) was conducted to uncover the underlying structure of the data, identify key patterns, and surface relationships that would guide subsequent modeling and business strategy. This stage combined statistical analysis with visual exploration to ensure a deep, intuitive understanding of customer and sales dynamics.

Correlation Analysis

To investigate relationships among critical numeric variables, a **correlation heatmap** was generated. This revealed several notable patterns:

- **Loyalty Score** demonstrated a strong positive correlation with **Customer Lifetime Value (CLV)**, indicating that customers deemed more loyal also tend to spend more over time.
- **Recency** (days since last purchase) showed a clear **negative correlation with Frequency and CLV**, underscoring that more recently active customers are typically frequent buyers and more profitable.
- Features like **Discount Utilization Rate** correlated modestly with Frequency, suggesting price sensitivity segments.

Time Series Patterns & Seasonality

Monthly sales trends were plotted to explore temporal dynamics. The analysis highlighted:

- **Distinct seasonal peaks**, often aligning with promotional periods, festivals, or year-end cycles.
- Notable dips during off-peak months, providing a basis for forecasting efforts and future campaign planning.

Decomposition of the time series using methods such as **STL decomposition** confirmed these seasonal effects alongside a stable underlying growth trend.

Demographic & Behavioral Patterns

By segmenting the data along demographic lines (Age, Income, Gender), the EDA uncovered several behavioral insights:

- Customers in the **25–35 age bracket** displayed higher transaction frequencies, likely driven by lifestyle or disposable income trends.
- Higher income groups generally corresponded with larger average transaction values and higher overall CLV.

- Gender-based analysis suggested roughly balanced spending, though product category preferences varied slightly.

Product & Discount Impact

Exploratory bar plots and boxplots by **Product Category** revealed differences in average transaction sizes and discount penetration. Certain categories were disproportionately reliant on discounting, suggesting opportunities for margin optimization.

Payment & Channel Preferences

Analysis of **payment methods and sales channels** indicated a growing preference for digital wallets and online sales platforms. This insight informs both marketing channel strategies and technical investments.

Key Takeaways from EDA

- **High Loyalty & CLV Clusters:** Correlation patterns justify focusing on loyalty metrics as leading indicators of profitability.
- **Seasonality Must Be Managed:** Identified peaks and troughs necessitate strategic stock, staffing, and promotional planning.
- **Targetable Demographics:** Younger, higher-income segments emerge as prime targets for growth and engagement campaigns.
- **Pricing & Discount Strategies:** Category-level discount analysis highlights where smarter pricing strategies could lift margins without harming volume.

Customer Segmentation

Identifying distinct customer segments is fundamental to deploying targeted marketing strategies and optimizing customer lifetime value. By leveraging advanced clustering techniques and enriching customer profiles with engineered features, we were able to segment the customer base into meaningful groups, each with distinct behaviors and business value.

Segmentation Approach

To move beyond simple heuristics, we employed sophisticated clustering algorithms:

- **Gaussian Mixture Models (GMM)**, which allowed for flexible, probabilistic assignments and captured elliptical cluster shapes in multi-dimensional space.
- **Agglomerative Hierarchical Clustering**, which created a dendrogram revealing how customers group together at various similarity thresholds, enabling validation and deeper exploratory insights.

The clustering was based on key behavioral features:

- **Recency**: Days since last purchase.
- **Frequency**: Number of purchases made.
- **Monetary Value (CLV)**: Total spend across the period.
- **Loyalty Score** and **Discount Utilization Rates** were also included to enrich the customer vectors.

Key Customer Segments Identified

The analysis surfaced four distinct segments:

1. **High-Value Loyalists**
    - Low recency (recent purchases), high frequency, and high CLV.
    - These customers are the core profit drivers and brand advocates.
2. **Mid-Tier Regulars**
    - Moderate frequency and monetary value, fairly regular purchase patterns.
    - Represent a stable revenue stream, ripe for upselling.
3. **Discount-Driven Deal Seekers**

- Transactions often coincide with promotions. Moderate frequency but sensitive to discounts.
- Benefit from personalized time-limited offers and bundled deals.

4. **Dormant or At-Risk Customers**
    - High recency (haven't purchased in a long time), low frequency and CLV.
    - Represent churn risks, candidates for reactivation campaigns.

Segment Profiles & Strategic Recommendations

Segment	**Characteristics**	**Recommended Strategy**
High-Value Loyalists	Repeat buyers, high spend & engagement	VIP loyalty programs, early product access
Mid-Tier Regulars	Consistent buyers, moderate spend	Targeted cross-sell & upsell campaigns
Discount-Driven Deal Seekers	Price sensitive, spike on promotions	Flash sales, personalized discount bundles
Dormant / At-Risk	Long inactivity, low spend	Win-back incentives, tailored outreach

Business Impact

This segmentation empowers marketing and CRM teams to design highly personalized campaigns, optimize promotion spending, and deepen relationships with each customer group. By prioritizing resources on high-value segments while strategically addressing churn risks, the business can maximize lifetime value and retention.

Sales Forecasting

Forecasting future sales is critical for planning marketing campaigns, managing inventory, staffing resources, and ensuring financial stability. By applying advanced time-series forecasting techniques to the historical transaction data, this analysis provides a data-driven outlook on upcoming sales trends, allowing the business to proactively strategize for demand fluctuations.

Forecasting Approach

The time-series modeling leveraged the strengths of the **Prophet algorithm**, developed by Facebook, which is highly effective for business datasets with clear seasonality and special event impacts. Prophet was selected because it:

- Automatically detects and models **weekly, monthly, and yearly seasonal patterns**.
- Handles **holidays and special events**, which are crucial in retail environments.
- Provides robust **prediction intervals**, offering insight into best and worst-case scenarios.

Data preprocessing included aggregating daily transaction values into monthly totals to focus on broader trends and minimize short-term noise. Additional time-based features such as month, quarter, and holiday flags were incorporated to further strengthen the model.

Forecasting Results

- The forecast revealed **clear seasonal peaks**, notably aligning with year-end festivals, local events, and promotional campaigns.
- Off-season dips were also identified, highlighting periods when sales naturally slow down, such as the first quarter post-holiday.
- Prediction intervals showed manageable uncertainty, instilling confidence in leveraging these forecasts for operational decisions.

A plotted forecast curve visually demonstrated projected sales for the next six months, with a band showing expected variation. This level of insight enables departments from marketing to logistics to synchronize their strategies.

Strategic Business Implications

- **Inventory & Supply Chain:** Adjust stock procurement and warehouse resources ahead of forecasted peaks to avoid stockouts and lost sales.
- **Marketing Campaigns:** Plan heavy promotional pushes during predicted upticks to capitalize on natural customer buying cycles.

- **Off-Peak Activation:** Introduce targeted offers or loyalty campaigns during expected slower periods to smooth revenue streams.
- **Financial Planning:** Incorporate forecasts into revenue projections and budget allocations, ensuring balanced cash flow management.

Why This Matters

Accurately forecasting sales does more than optimize operations—it reduces excess inventory costs, improves customer satisfaction through product availability, and allows for smarter resource allocation across the business.

Churn Prediction

Customer churn poses a direct threat to long-term revenue and profitability. Identifying customers at risk of leaving allows the business to proactively intervene and implement retention measures that protect the lifetime value of its customer base. This analysis deployed advanced machine learning techniques to predict churn with high reliability.

Modeling Approach

To predict churn, we constructed a supervised classification model using **XGBoost (Extreme Gradient Boosting)**, a powerful algorithm well-suited for tabular data with heterogeneous features. The model incorporated a range of behavioral and transactional indicators, including:

- **Recency:** Time since the last purchase.
- **Frequency:** Total number of purchases.
- **Monetary Value (CLV):** Lifetime spend.
- **Loyalty Score** and **Discount Utilization Rate.**

The dataset was balanced through sampling to mitigate bias toward the majority non-churn class. Hyperparameter tuning was executed via **GridSearchCV**, optimizing for metrics such as ROC AUC to maximize separability.

Model Performance

- The final XGBoost model achieved strong performance, with **high precision and recall**, minimizing both false positives and false negatives.
- The **ROC AUC curve** indicated excellent discriminative capability, confirming the model's reliability in distinguishing churn-prone customers from loyal ones.

A confusion matrix analysis reinforced confidence that the model could accurately flag at-risk customers without overwhelming marketing teams with unnecessary retention efforts.

Strategic Business Recommendations

- **Targeted Retention Campaigns:** Utilize churn scores to prioritize outreach. For example, offer personalized incentives or loyalty bonuses to customers identified as high risk.
- **Customer Experience Interventions:** Proactively engage predicted churners with satisfaction surveys, exclusive previews, or direct account management contact.
- **Lifecycle Automation:** Implement automated marketing workflows that trigger based on churn risk levels, ensuring timely intervention.

Business Value

Effective churn prediction translates directly to preserved revenue. Retaining even a fraction of at-risk customers identified by the model safeguards significant lifetime value, reduces acquisition costs needed to replace lost customers, and sustains a stronger brand reputation through consistent engagement.

Market Basket Analysis

Understanding which products customers commonly purchase together enables businesses to craft compelling cross-sell offers, optimize store layouts (online or offline), and boost overall basket sizes. Market Basket Analysis, leveraging Association Rule Mining, reveals these product affinities hidden within transaction data.

Analytical Approach

The **Apriori algorithm** was employed to identify frequent itemsets within historical transaction records. From these, strong association rules were derived based on:

- **Support:** The proportion of transactions containing a specific itemset.
- **Confidence:** The likelihood of purchasing item B given item A is purchased.
- **Lift:** How much more likely item B is purchased with item A versus being purchased independently.

By setting minimum thresholds on support and confidence, the analysis focused on meaningful, actionable associations, avoiding spurious or low-impact combinations.

Key Insights & Discovered Rules

- **Complementary Pairings:** Rules highlighted frequent combinations such as complementary electronics and warranty packages, or apparel and accessory bundles.
- **Upsell Opportunities:** Customers buying mid-tier products often showed elevated lift toward premium add-ons, indicating a natural propensity to upgrade when presented effectively.
- **Category Cross-Links:** Purchases in home essentials were commonly linked to seasonal decor, suggesting opportunities for seasonal campaigns that bundle these categories.

Business Implications

- **Strategic Product Bundling:** Introduce curated product bundles based on top association rules to increase average transaction value. For instance, bundling kitchen appliances with popular maintenance kits.
- **Personalized Cross-Sell Promotions:** Use association insights to design personalized recommendations. A customer purchasing a smartphone could automatically receive offers on compatible cases or insurance plans.

- **Inventory Planning:** Recognize products with strong cross-sell links and ensure they are stocked together, reducing friction in both online shopping carts and physical store layouts.

Enhancing Customer Experience

By surfacing intuitive combinations, market basket analysis not only grows sales but also enhances the customer experience—making it easier for customers to discover items they're likely to need, while increasing convenience and perceived value.

Strategic Recommendations

Based on the comprehensive analysis of customer demographics, transaction patterns, segmentation, forecasting insights, churn prediction, and market basket affinities, the following strategic recommendations are proposed to drive sustained business growth, optimize marketing investments, and maximize customer lifetime value.

1. Prioritize High-Value Customer Segments

The segmentation analysis identified a core group of **high-value loyalists** with low recency, high frequency, and substantial lifetime spend. To deepen engagement with this segment:

- **Launch VIP programs** offering exclusive early access to new products, personalized services, and tiered loyalty benefits.
- Invite these customers to participate in feedback initiatives, strengthening emotional loyalty while gathering valuable insights.

2. Proactively Manage Churn Risk

The churn prediction model highlighted customers at elevated risk of attrition. To address this:

- Deploy **targeted retention campaigns**, such as personalized discounts, loyalty top-ups, or win-back emails.
- Integrate churn risk scores into CRM workflows, enabling customer service teams to prioritize high-risk contacts with proactive outreach.

3. Leverage Cross-Selling and Bundling Opportunities

Market basket analysis revealed strong product associations that can be harnessed to increase average basket sizes:

- **Bundle frequently co-purchased items** into convenience packs or themed kits, offering slight bundled discounts to drive uptake.
- Use personalized recommendations on e-commerce platforms, based on prior purchase patterns and discovered association rules.

4. Align Operations with Forecasted Demand

The Prophet-based sales forecasts identified clear seasonal peaks and troughs:

- **Scale inventory and staffing levels** in anticipation of projected demand spikes to prevent stockouts and maintain service quality.

- Introduce off-season incentives or limited-time offers to smooth demand during slower periods, stabilizing cash flows.

5. Optimize Pricing and Discount Strategies

The EDA and segmentation highlighted customer segments that are highly discount-sensitive:

- Strategically apply discounts to price-sensitive segments (e.g., deal seekers), while maintaining regular pricing for high-loyalty customers less driven by promotions.
- Test and refine promotional cadences to protect margins without sacrificing volume.

6. Enhance Digital & Payment Channel Experience

Given the observed customer tilt toward digital wallets and online purchases:

- Invest in seamless, mobile-optimized shopping experiences.
- Offer flexible, secure payment options to capture evolving customer preferences.

Business Impact

Collectively, these strategies position the business to enhance customer engagement, safeguard against churn, grow average order values, and align operational tactics with predictive insights—creating a robust foundation for sustainable, data-driven growth.

Conclusion

This report provided a rigorous, multi-dimensional analysis of customer and sales data to uncover actionable insights that can directly inform strategic business decisions. By integrating data preprocessing, advanced feature engineering, exploratory analytics, segmentation, forecasting, churn prediction, and market basket analysis, the study delivers a holistic view of customer behaviors and revenue drivers.

The journey began with careful data cleaning and enrichment, transforming raw transaction records into a powerful analytical asset. Exploratory Data Analysis revealed important patterns—such as the strong link between loyalty and lifetime value, clear seasonality in sales trends, and demographic nuances influencing purchasing behavior.

Customer segmentation, powered by Gaussian Mixture Models and hierarchical clustering, unveiled distinct profiles ranging from high-value loyalists to discount-sensitive deal seekers and dormant customers. Each segment presents unique opportunities for tailored marketing, enabling more precise allocation of promotional budgets and stronger customer relationships.

Time-series forecasting using the Prophet model produced clear visibility into upcoming sales dynamics, empowering the business to prepare inventory, workforce, and campaigns in line with anticipated demand. Meanwhile, predictive churn modeling with XGBoost equipped the organization with a powerful tool to flag at-risk customers, opening the door for timely, personalized retention efforts.

Market Basket Analysis rounded out the insights by uncovering frequently co-purchased products, laying the groundwork for strategic bundling and cross-sell initiatives that can drive higher basket values and enhance the customer experience.

Looking Forward

The findings of this analysis translate directly into strategic levers that the business can pull to optimize revenue growth, deepen customer loyalty, and safeguard future profitability. By:

- Prioritizing high-value customers with exclusive engagement tactics,
- Proactively addressing churn risks before revenue erosion occurs,
- Designing smart cross-sell bundles based on actual purchase affinities,
- And aligning operational plans with data-backed forecasts,

the company positions itself to move beyond reactive decision-making and embrace a culture of proactive, data-driven strategy.

Final Thought

Data is more than a record of past transactions; it is a blueprint for future success. By continuing to invest in advanced analytics and leveraging insights across departments, the organization will build enduring competitive advantages, ensuring it not only meets customer expectations—but consistently exceeds them.