Info Bharat Interns

Internship Project on Customer Sales and Sales Data Analysis

by S. Sushik Nanda

1.Project Overview:

This project is a comprehensive customer analytics and forecasting pipeline designed to deliver actionable business insights from raw transactional data. The goal was to extract, clean, and analyze customer purchase behavior, uncover patterns in sales performance, segment customers, predict churn, and forecast future sales, ultimately supporting data-driven decision-making for strategic growth.

The workflow begins with data preprocessing and feature engineering, where missing values are addressed using advanced imputation techniques like KNN and MICE. Outliers are treated with Tukey's method, and relevant variables such as Customer Lifetime Value (CLV) and total sales are engineered. Scaling is applied to ensure model robustness.

Following this, Exploratory Data Analysis (EDA) explores key trends and relationships using interactive visualizations. Sales are broken down by time periods, categories, demographics, and discounts. Seasonal decomposition (STL) is used to uncover sales cycles, and correlation matrices highlight influential variables.

Next, the project employs customer segmentation using RFM (Recency, Frequency, Monetary) analysis combined with clustering algorithms like Gaussian Mixture Models and Agglomerative Clustering. These segments are profiled in detail, supporting tailored marketing strategies such as loyalty rewards, win-back campaigns, or targeted promotions.

The sales forecasting section integrates Prophet, SARIMA, and LSTM models to capture trend, seasonality, and sequential patterns. Forecasts help the business prepare for expected dips and peaks, informing promotional timing and inventory planning. Model performance is evaluated with MAE, RMSE, and MAPE metrics.

For churn prediction, machine learning models—including Random Forest, Logistic Regression, Decision Tree, and XGBoost—identify customers at risk of leaving. Feature importance is assessed, and models are optimized with GridSearchCV and evaluated using classification metrics and precision-recall curves.

Finally, market basket analysis uncovers frequently co-purchased products using the FP-Growth algorithm. Strong association rules, measured by lift and confidence, provide insights into bundling and cross-sell opportunities. Profitability analysis highlights which products may benefit from discount adjustments or promotions.

This project delivers a full analytics solution—from raw data to strategic insights—supporting customer retention, profit optimization, and sustainable sales growth through data science.

2. Data Preprocessing and Feature Engineering:

In this project, data preprocessing and feature engineering were foundational steps that ensured clean, consistent, and insightful data for advanced analytics. The process began by

loading the raw dataset and identifying key numerical columns, including Age, Income Level, Loyalty Score, Cost Price, and Selling Price. Since real-world datasets often contain missing values or inconsistencies, missing values in these fields were handled through imputation. Depending on the configuration, the code supported both KNN Imputer and MICE (Iterative Imputer) techniques to fill gaps intelligently, maintaining data integrity.

To further improve data quality, non-numeric entries were coerced to NaN and managed accordingly. Columns that had all missing values were either skipped or filled with a default value (like zero or the median). Afterward, Tukey's method was applied to cap outliers across numeric variables. This technique effectively reduced the influence of extreme values that could distort machine learning models or statistical analysis.

Next, the data was normalized using either StandardScaler or MinMaxScaler (based on the chosen flag), ensuring that variables were on a consistent scale—especially important for distance-based models and clustering algorithms.

Beyond cleaning, the code constructed meaningful new features. Two critical engineered metrics were:

- Total Sale, calculated as Unit Price × Purchase Quantity, representing transaction-level revenue.
- Customer Lifetime Value (CLV), derived by multiplying Total Sale × Loyalty Score, which serves as a proxy for long-term customer worth.

Advanced customer-level attributes like Recency (days since last purchase), Frequency (total number of purchases), and other behavioral metrics were also computed. These were essential inputs for customer segmentation and churn modeling later in the pipeline.

By the end of this phase, the dataset was not only cleaned and consistent but also enriched with powerful new features that reflect customer behavior, sales performance, and profitability trends. This rigorous preprocessing laid a strong foundation for robust analytics, enabling accurate segmentation, forecasting, and predictive modeling.

3. Exploratory Data Analysis (EDA):

The Exploratory Data Analysis (EDA) phase was crucial in uncovering meaningful patterns and trends in the dataset, providing a strong foundation for deeper analysis and modeling. A variety of visual and statistical techniques were applied to understand customer behavior, sales dynamics, and overall business performance.

To begin with, sales trends across product categories were examined using interactive bar charts. This revealed which categories generated the most revenue and helped pinpoint top-performing segments. Furthermore, by analyzing transactions by day of the week, the team identified peak purchasing days, offering useful insights for scheduling promotions and staffing.

A seasonal decomposition of monthly sales data using STL (Seasonal-Trend decomposition using LOESS) highlighted recurring patterns and seasonality within the business cycle. This was complemented by time-of-day analysis, which illustrated customer purchase behavior across different hours, aiding in understanding demand flow.

In terms of customer demographics, visualizations of sales across age groups showed that certain age bands contributed more significantly to revenue. Gender-based purchasing patterns and income-level-based sales distributions added further context to the customer profile, helping in segmentation and personalized marketing.

The impact of discounts on total sales was assessed via scatter plots and correlation coefficients. While discounts generally led to increased sales, the analysis also raised considerations around margin trade-offs. A correlation heatmap was constructed to explore relationships between key numerical features like recency, frequency, total sale, loyalty score, and discount offered. This matrix provided insights into how variables interacted, offering clues for future modeling.

Finally, the EDA extended to payment method preferences. Where available, the data revealed the most popular transaction methods among customers, informing operational and marketing decisions.

Overall, the EDA phase illuminated not just what customers were buying, but when, how, and under what conditions. These findings laid the groundwork for more advanced tasks such as segmentation, churn modeling, and forecasting — making EDA not just an exploratory phase, but a strategic one.

4.Customer Segmentation:

To better understand and serve different types of customers, a comprehensive segmentation approach was adopted using both RFM (Recency, Frequency, Monetary) analysis and advanced clustering techniques.

First, each customer was evaluated on three core behavioral metrics:

- Recency: How recently they made a purchase
- Frequency: How often they purchased
- Monetary: The total value of their purchases

This data was normalized and used as input for two clustering methods: Gaussian Mixture Models (GMM) and Agglomerative Clustering. GMM allowed for soft probabilistic clustering, providing nuanced distinctions between customer types, while Agglomerative Clustering grouped customers based on hierarchical similarity.

The segmentation process revealed four distinct customer segments, each with unique traits:

- Segment 0: Comprised of loyal, high-frequency shoppers with strong purchasing power. These individuals often return and exhibit high Customer Lifetime Value (CLV). Targeting them with early-access sales and exclusive rewards can enhance retention.
- Segment 1: Represented customers who haven't engaged recently. Their declining activity indicates a need for win-back campaigns or reactivation offers.
- Segment 2: Though recent in activity, this group purchases infrequently and spends less. They're ideal candidates for upselling strategies through personalized discounts or bundles.
- Segment 3: Moderately active customers who engage occasionally and show average monetary value. Regular promotions and engagement campaigns may help nurture their value.

To refine the segmentation, the analysis incorporated additional dimensions like Loyalty Score, Discount Utilization, and Payment Method Preferences, providing a more holistic view of customer behaviors and preferences.

The outcome of this segmentation enables targeted marketing strategies, tailored product offerings, and smarter loyalty programs. By aligning communications and promotions with the specific characteristics of each segment, the business can drive higher engagement, increase sales, and improve customer satisfaction.

This customer-centric strategy transforms raw behavioral data into actionable insights, laying the foundation for sustainable growth and improved customer relationships.

5.Sales Forecasting:

In this project, sales forecasting was approached through a comprehensive time series modeling pipeline, leveraging three distinct forecasting techniques—Prophet, SARIMA, and LSTM—each chosen for its strengths in capturing unique patterns in sales behavior.

The first model employed was Facebook Prophet, a robust tool known for handling time series data with multiple seasonality patterns. It utilized historical daily sales data to project future trends over a 90-day period. This model helped visualize potential spikes and slumps, enabling the business to proactively plan around expected fluctuations. Prophet's intuitive framework allowed for rapid iteration and clear visual outputs.

Complementing this, the SARIMA (Seasonal AutoRegressive Integrated Moving Average) model was used to capture both trend and seasonality in the monthly aggregated sales data. It effectively modeled historical patterns while offering more traditional statistical insight into cyclical sales behavior. Model evaluation using MAE, RMSE, and MAPE revealed a reliable performance, with SARIMA pinpointing both a potential dip and a peak in sales within the next six months. These findings are invaluable for inventory planning and resource allocation during high- and low-demand periods.

The third and more advanced method involved a Long Short-Term Memory (LSTM) neural network. By training on scaled, sequential monthly sales data, the LSTM captured complex temporal dependencies often missed by statistical models. While more computationally intensive, this deep learning approach provided a nuanced forecast trajectory that aligned well with actual test data.

Each model was assessed for performance, and together they painted a more complete picture of future sales. Notably, all models indicated a potential slowdown during off-season months—signaling the need for promotional efforts to sustain momentum. Meanwhile, anticipated peaks suggested the importance of ramping up marketing and inventory efforts in advance.

In summary, this multi-model strategy not only ensured forecasting robustness but also provided actionable insights for the business. By triangulating results from Prophet, SARIMA, and LSTM, stakeholders can better align operations with market demand, optimize promotional timing, and mitigate risks tied to sales volatility.

6.Predictive Modeling: Customer Churn:

Understanding and preventing customer churn is crucial for sustaining growth and profitability. In this project, a predictive churn model was developed to identify customers who are at risk of discontinuing their engagement with the business. This enables timely interventions and targeted retention strategies.

To begin, a binary churn variable was generated, simulating whether a customer has churned or not. Key features that influence churn — such as customer Age, Loyalty Score, Recency (how recently a purchase was made), Frequency (number of purchases), and Discounts Offered — were selected for modeling.

Several machine learning algorithms were trained and evaluated. These included Random Forest, Logistic Regression, Decision Trees, and XGBoost. The dataset was split into training and testing sets to ensure unbiased evaluation. Among these models, Random Forest and XGBoost stood out with their ability to handle feature importance, non-linearity, and high performance in classification tasks.

To further enhance accuracy, GridSearchCV was used to fine-tune the Random Forest model. The best combination of parameters was selected based on AUC-ROC, a metric that balances sensitivity and specificity.

The models were assessed using various evaluation metrics such as precision, recall, F1-score, and confusion matrices. A precision-recall curve was also plotted, offering a visual insight into the model's capability to correctly identify churners. This helped in selecting the right threshold for intervention strategies.

Key insights from the model showed that customers with high recency (i.e., they haven't purchased recently), low loyalty scores, and declining purchase frequency are more likely to

churn. Discounts were also a contributing factor — some customers responded well to them, while others didn't convert despite offers.

Overall, this churn prediction model provides a solid foundation for implementing proactive customer retention plans. By targeting high-risk individuals early — especially those who show declining engagement — businesses can reduce churn rates, improve customer lifetime value (CLV), and strengthen long-term relationships.

7.Product Analysis and Cross-Selling Strategy:

The product analysis and cross-selling strategy conducted in this project provides crucial insights for boosting sales and enhancing customer experience. Using advanced Market Basket Analysis techniques like the FP-Growth algorithm, frequently co-purchased product combinations were identified. These associations, evaluated using metrics such as lift and confidence, revealed high-performing item pairs that can be bundled to encourage cross-selling. For instance, sub-category combinations with strong co-occurrence can be promoted through combo deals or “frequently bought together” campaigns, increasing both average order value and customer satisfaction.

On the profitability side, the analysis showed clear distinctions between products generating high net profit and those consistently underperforming due to heavy discounting. By calculating profit per unit and adjusting for discounts, a list of low-margin products was flagged as prime candidates for promotional bundles or price strategy review. This data-driven approach ensures promotions are not just driving volume but are also aligned with profitability goals.

Altogether, this strategic blend of product affinity analysis and profit-based targeting empowers businesses to craft smarter marketing campaigns. The recommendations include bundling high-affinity products and revisiting discount strategies for low-profit items — helping to optimize both customer value and revenue potential.

8.Reporting and Visualization:

The reporting and visualization phase played a vital role in transforming raw data into meaningful business insights. A variety of visual tools were used to make sense of customer behavior, product performance, and sales dynamics. Bar charts and time-series plots helped break down sales trends by category, day of the week, and hour of the day, offering a clear picture of when and what customers tend to purchase most.

Demographic-based visualizations—such as age group vs. sales, gender-based purchase patterns, and income-level comparisons—enabled a deeper understanding of the customer base. Correlation heatmaps were used to uncover relationships between variables like discount offered, total sales, and loyalty score.

Advanced techniques like STL decomposition highlighted seasonality in monthly sales, while scatter plots illustrated how discounts influenced revenue generation. The analysis of payment method preferences also added a practical layer, offering insights into customer convenience.

Altogether, these visualizations weren't just for aesthetic presentation—they directly supported strategic decisions in segmentation, forecasting, churn prediction, and product bundling. They made the data story intuitive for business stakeholders, turning complex metrics into actionable recommendations. This visualization-driven reporting created a bridge between technical analysis and business strategy, making insights both accessible and impactful.