Info Bharat Interns

Internship Project on Customer Sales and Sales Data Analysis

by SRI KRISHNA KODURI

1.Project Overview:

This project presents a comprehensive data-driven analysis pipeline built around a retail dataset to extract meaningful business insights, improve decision-making, and enable predictive capabilities. It begins with data preprocessing, where missing values are imputed, outliers are handled, and features are standardized. Advanced feature engineering techniques like RFM (Recency, Frequency, Monetary) analysis, Customer Lifetime Value (CLV) estimation, and customer loyalty scoring are employed to enrich the dataset.

The project proceeds with exploratory data analysis (EDA) using visualizations to understand customer behavior and sales trends. For time-series forecasting, three models—SARIMA, Prophet, and LSTM—are used to predict future sales, comparing accuracy through RMSE, MAE, and MAPE metrics. To enhance customer understanding, clustering techniques like Gaussian Mixture Models and Agglomerative Clustering segment customers into meaningful groups.

Additionally, machine learning models including Random Forest, Logistic Regression, and XGBoost are trained for churn prediction, helping to identify at-risk customers. Finally, Market Basket Analysis using the Apriori algorithm uncovers product bundling and cross-selling opportunities.

This end-to-end pipeline not only demonstrates solid data science practices but also simulates real-world retail analytics scenarios, making it a valuable exercise in combining statistical modeling, machine learning, and business strategy.

2. Data Preprocessing and Feature Engineering:

Effective data preprocessing and feature engineering form the foundation of any robust data science pipeline, especially in domains like retail analytics where raw data can often be messy, incomplete, or inconsistent. In this project, both these steps were handled meticulously to ensure the dataset was refined and enhanced for deeper analysis and modeling.

Data Preprocessing

The preprocessing began with handling the Date of Purchase column, converting it to a datetime format. This step was essential to facilitate any time-series analysis, such as forecasting sales or measuring customer recency.

To tackle missing values in numerical columns, the K-Nearest Neighbors Imputer (KNNImputer) was employed. This method estimates missing values by averaging those of neighboring data points, making it more robust than simply filling with mean or median values. An alternative, IterativeImputer, was also considered for more complex imputation scenarios but left commented for flexibility.

Next, outlier treatment was performed using Tukey's method, which caps extreme values based on interquartile range thresholds. Additionally, an advanced alternative—Robust Z-Score (RZS)—was optionally included. This method, based on median and median absolute deviation (MAD), offers better resistance to skewed distributions compared to traditional Z-scores.

Once the data was cleaned, standardization using StandardScaler was applied to the numerical columns. This transformed the values to have zero mean and unit variance, which is particularly beneficial when features have varying units and scales. As an optional alternative, the code also included provisions for MinMaxScaler, which rescales features into a [0, 1] range for algorithms sensitive to magnitudes.

Feature Engineering

With a clean dataset in place, the focus shifted to deriving meaningful features that could capture customer behavior and enhance model performance. One of the most valuable outputs was the construction of RFM features — *Recency*, *Frequency*, and *Monetary* values:

- Recency was calculated as the number of days since the customer's most recent purchase.
- Frequency reflected the count of unique purchases per customer.
- Monetary measured the total amount spent.

Using these three components, a base RFM DataFrame was created, forming the basis for further enhancements.

To dive deeper into customer value, Customer Lifetime Value (CLV) was approximated. This was calculated using a simplified formula: Average Order Value × Purchase Frequency × Profit Margin. A 10% profit margin was assumed for this purpose, providing a reasonable estimate of long-term revenue potential per customer.

The Average Purchase Frequency metric was also introduced. It normalized customer activity by dividing their purchase count by the total time span (in months) between their first and last purchase. This helped distinguish between customers who made frequent purchases in short bursts versus those who bought sporadically over a long period.

Further, the code explored Discount Utilization Rate, assuming relevant columns (Discount Amount and Selling Price) were present. This metric shows how much discount a customer typically avails, helping segment price-sensitive buyers. Similarly, Payment Method Preference was derived using the mode of each customer's transaction history, offering insights into customer habits and operational trends.

A crucial step was the construction of a Loyalty Score, based on quantiles of the Frequency metric. Customers were grouped into four tiers (1 to 4) using qcut, which assists in clustering and targeting efforts.

To top it off, engineered metrics were reused to label churn probability, forecast sales, segment customers, and even run association rule mining. This multi-purpose reusability highlights how strong feature engineering accelerates downstream tasks.

In summary, the data preprocessing and feature engineering sections of your pipeline were thorough and thoughtfully implemented. From correcting missing data and handling outliers to building advanced metrics like CLV and discount utilization, every step ensured the dataset was ready for high-quality insights and predictive modeling. These foundational stages laid the groundwork for customer segmentation, forecasting, and strategic business recommendations that followed in the later stages of your notebook.

3. Exploratory Data Analysis (EDA):

The Exploratory Data Analysis (EDA) phase played a crucial role in understanding the retail dataset's underlying structure, trends, and customer behavior. It began with converting the 'Date of Purchase' column into a datetime format to enable time-based analysis. Missing values in numerical columns were addressed using the K-Nearest Neighbors (KNN) imputer, ensuring no gaps in critical features like purchase amount or quantity.

To manage skewness and reduce the influence of extreme values, outliers were handled using Tukey's method and the Robust Z-Score approach, capping values that fell beyond statistically acceptable ranges. Following this, all numerical features were standardized using StandardScaler, bringing them to a uniform scale for better visualization and machine learning readiness.

Feature engineering added depth to the analysis. RFM metrics—Recency, Frequency, and Monetary value—were computed to profile customers based on their shopping habits. These insights were enriched with calculated Customer Lifetime Value (CLV), average purchase frequency, discount utilization rate, and preferred payment methods. These features provided a more holistic picture of consumer patterns.

Visualization played a key role. A histogram illustrated the distribution of customer recency, while time series plots showed monthly sales trends and seasonal patterns. STL decomposition further revealed underlying seasonality and trend components in sales. A correlation heatmap gave insights into how numerical variables relate to each other, supporting deeper analysis and future model decisions.

In summary, the EDA process effectively cleaned and transformed the dataset, created meaningful customer-centric metrics, and visualized key patterns in behavior and time trends. These steps laid a strong foundation for segmentation, forecasting, and predictive modeling tasks that followed.

4.Customer Segmentation:

To gain deeper insights into consumer behavior, a comprehensive customer segmentation was conducted using clustering techniques. The foundation of this process was built upon RFM metrics — Recency, Frequency, and Monetary — along with the Loyalty Score, all of which capture how recently and frequently a customer purchases and how much they spend.

The Gaussian Mixture Model (GMM), a probabilistic clustering technique, was applied to classify customers into four distinct segments. This model allowed for soft clustering, meaning customers had a probability of belonging to each cluster, resulting in more nuanced groupings.

Once segmentation was done, visualization using pair plots revealed clear distinctions between groups based on spending patterns and activity levels. For example, high-frequency and high-monetary customers were clustered separately from low-value or inactive ones. This enables targeted strategies — such as loyalty programs for high-value customers and reactivation campaigns for dormant ones.

To enrich the segmentation further, features such as Customer Lifetime Value (CLV), Average Discount Utilization, and Preferred Payment Method were included where available. These behavioral metrics provided a 360-degree view of each segment's purchasing style and financial impact.

Moreover, segment profiling highlighted the average characteristics of each group, including their spending habits and loyalty scores. For instance, one segment demonstrated frequent purchases with high monetary value and low recency — indicating loyal, high-value customers.

In summary, this segmentation lays the groundwork for personalized marketing, efficient resource allocation, and strategic decision-making. With this approach, businesses can focus their efforts where they matter most — increasing profitability and customer satisfaction through data-driven actions.

5.Sales Forecasting:

In the sales forecasting section, multiple advanced models were applied to predict future sales trends and guide strategic decisions. Initially, the dataset was processed to generate monthly sales figures, which were then modeled using SARIMA, Prophet, and LSTM approaches.

The SARIMA model effectively captured seasonality, showing strong performance during regular sales cycles and offering relatively low error metrics. Prophet, enhanced with holiday indicators and custom seasonality, provided intuitive and visually interpretable forecasts, particularly useful around festivals or promotions. LSTM, a deep learning model, handled

nonlinear trends well and offered flexibility in capturing complex time-based patterns—though it required more computational resources and careful tuning.

Each model was evaluated using metrics such as MAE, RMSE, and MAPE, allowing for a performance comparison. The results consistently revealed sales peaks around major holidays (like December and October), while off-season dips (April–June) highlighted opportunities for targeted promotions or clearance sales.

The overall analysis enables data-driven planning: ramping up inventory and marketing during high-demand periods and deploying offers during slumps. This multi-model forecasting approach empowers the business with both short-term insights and long-term strategic foresight, ensuring well-informed and proactive decisions.

Top of Form

6.Predictive Modeling: Customer Churn:

In this analysis, customer churn was predicted using machine learning models by analyzing past purchase behavior. Churn was defined as a customer not making any purchase within the last 90 days. The dataset was enriched with key features such as Recency, Frequency, Monetary value, and Loyalty Score, which were then used to train multiple models.

The Random Forest classifier was initially trained to classify churners, achieving a good balance between precision and recall. To further strengthen the prediction pipeline, other models including Logistic Regression, Decision Tree, and XGBoost were also evaluated. Each model's performance was assessed using confusion matrices, classification reports, and ROC-AUC scores, providing a clear understanding of their accuracy and effectiveness.

Among them, XGBoost and Random Forest showed promising results in identifying at-risk customers. The ROC and precision-recall curves provided additional validation by visually demonstrating model capability.

This churn prediction framework can be valuable for businesses to proactively engage customers who are likely to leave, using targeted campaigns or personalized offers. Ultimately, it helps in reducing customer loss and improving long-term revenue by shifting from reactive to preventive strategies.

7.Product Analysis and Cross-Selling Strategy:

To enhance revenue through smarter recommendations, Market Basket Analysis was implemented using the Apriori algorithm. By analyzing the product combinations in customer invoices, frequently co-purchased items were identified. These associations provide clear

insights into which products naturally bundle together, enabling the business to suggest relevant cross-sell or upsell items at checkout or in marketing campaigns.

The transactional data was converted into a basket matrix, marking whether each product was purchased per invoice. Association rules were extracted using metrics such as lift and confidence. High-confidence rules indicated strong purchase patterns—for example, if a customer buys Product A, there's a significant likelihood they'll also purchase Product B. This kind of insight supports curated product bundles, combo offers, and personalized promotions.

Furthermore, these rules help inform shelf placement, combo discounts, and targeted online recommendations, ultimately boosting conversion rates and average order value. For instance, if two items are often bought together, positioning them side-by-side—physically or digitally—can nudge users to buy both.

In summary, this analysis not only highlights popular product pairings but also unlocks actionable strategies for maximizing customer value through cross-selling—delivering both better shopping experiences and improved business performance.

8.Reporting and Visualization:

In the final phase of the project, a variety of reporting and visualization techniques were used to interpret and communicate the insights gained from data analysis. Visual tools like seaborn, matplotlib, and plotly helped illustrate key business metrics such as sales trends, customer segments, and churn probabilities. For instance, monthly sales were visualized to spot seasonal spikes and dips, aiding in planning promotions and inventory.

Advanced forecasting models like SARIMA, Prophet, and LSTM were not only evaluated using RMSE, MAE, and MAPE, but also presented visually to contrast predictions with actual data. Prophet's customizable seasonal plots gave clarity on how festivals and time-based patterns affect sales. Furthermore, clustering results from customer segmentation were presented using pair plots to visually distinguish different customer groups based on RFM (Recency, Frequency, Monetary) and loyalty scores.

Additionally, ROC and Precision-Recall curves were plotted for churn prediction models (Random Forest, XGBoost, etc.), helping stakeholders understand model performance beyond mere accuracy. These visual summaries made it easier for non-technical stakeholders to grasp the business impact of the models and make data-driven decisions confidently. Overall, reporting and visualization acted as the bridge between raw analysis and actionable strategy.

Top of Form

Bottom of Form

Bottom of Form