**Retail Store & Customer Insights: An MLOps Capstone Project**

**Submitted by:** Santhosh.v

**1. Executive Summary**

This report details the development of an end-to-end analytics and machine learning pipeline for a retail chain. The project's primary goal was to leverage daily sales data to understand store performance and customer behavior. The pipeline successfully processed raw data, engineered key features, built a customer segmentation model, and deployed actionable insights via a FastAPI-powered API. The key findings include the identification of top-performing stores, an understanding of seasonal sales trends, and the classification of customers into high-value and loyal segments for targeted marketing.

**2. Project Objective**

The retail chain aims to better understand store performance and customer behavior to steer targeted marketing campaigns and inventory strategy. The project's objective was to build a comprehensive analytics pipeline to:

* Process and analyze daily sales data across stores and regions.
* Perform RFM-based customer loyalty analysis.
* Build machine learning models for customer segmentation.
* Deploy insights as APIs for business dashboards.
* Surface actionable insights on top-performing stores, loyal customers, and seasonal patterns for business decision-making.

**3. Data Processing & Analysis**

The project began with the ingestion and processing of the provided customer shopping dataset.

* **Data Ingestion & Cleaning**: The raw data was loaded, and the invoice\_date column was cleaned and formatted to the correct datetime format. No null values were present, which simplified the cleaning process.
* **Feature Engineering**: New features were created to support the analysis:
  + **Total Price**: Calculated as quantity \* price to represent the total revenue for each transaction. This serves as a proxy for profitability, as no discount data was available in the dataset.
  + **Time-Series Features**: invoice\_year, invoice\_month, and invoice\_quarter were extracted from the invoice date for seasonal trend analysis.
* **Key Findings**:
  + **Store Performance**: Total revenue was analyzed across all shopping\_mall locations to identify the top-performing stores.
  + **Seasonal Trends**: A time-series analysis of monthly sales revealed potential seasonal patterns in customer purchasing behavior.
  + **Payment Method**: The distribution of payment\_method showed the most common ways customers pay.

**4. Customer Segmentation**

To understand customer behavior, an RFM-based segmentation model was developed.

* **Methodology**:
  + **RFM Calculation**: Recency, Frequency, and Monetary scores were computed for each customer\_id.
  + **Clustering**: The K-Means clustering algorithm was applied to the scaled RFM data. The Elbow Method was used to determine the optimal number of clusters for segmentation.
* **Cluster Profiles**: The model identified three distinct customer segments:
  + **High-Value/Loyal Customers**: Characterized by high frequency and high monetary value. These are the most valuable customers who should be targeted for loyalty programs.
  + **Recent/Frequent Customers**: These customers shop frequently but have a lower monetary value. They are ideal for up-selling and cross-selling campaigns.
  + **New/One-Time Customers**: Defined by high recency but low frequency and monetary value. The goal for this segment is to convert them into repeat customers.

**5. MLOps Pipeline & Deployment**

The final phase involved operationalizing the insights through an API and establishing a framework for MLOps best practices.

* **API Development**: A FastAPI application was built to deploy the insights as APIs for business dashboards. Key endpoints were created to serve data on store performance, top customers, sales trends, and customer segments.
* **ML Model Deployment**: The trained K-Means model was integrated into the API to classify new customers in real-time.
* **Continuous Improvement**: Although not physically built, a conceptual framework for a CI/CD pipeline, monitoring, and automated model retraining was established to ensure the system remains reliable and the model's performance doesn't degrade over time.

**6. Conclusion**

The project successfully delivered a robust analytics and machine learning pipeline that provides clear, actionable insights for the retail chain. By segmenting customers and analyzing store performance, the business can now implement data-driven strategies for marketing, inventory management, and customer relationship management. This project serves as a strong foundation for a scalable and maintainable MLOps system.