Project Report: Retail Analytics and Forecasting

Executive Summary

This report summarizes the retail analytics and forecasting project based on the provided Jupyter Notebook ("Retail_Analytics_and_Forecasting.ipynb"). The project involves loading and analyzing a retail sales dataset, performing exploratory data analysis (EDA), customer/store segmentation using clustering, and sales forecasting using time series models. Key insights include seasonal sales trends, product category performance, store benchmarks, and recommendations for optimization. The analysis uses Python libraries such as Pandas, NumPy, Matplotlib, Seaborn, Scikit-learn (for KMeans clustering), and Statsmodels (for SARIMAX forecasting).

The dataset covers sales transactions from 2023 to 2024 across various stores in India, including details on products, pricing, discounts, and customer behavior. The project aims to provide data-driven insights for retail management, enabling better inventory planning, pricing strategies, and demand forecasting.

Introduction

Project Objective

The goal of this project is to:

- Analyze retail sales data for trends, patterns, and performance metrics.

- Segment stores or customers using clustering techniques to identify high/low performers.

- Forecast future sales using time series modeling to support proactive decision-making.

- Derive actionable business insights and recommendations.

Scope

- Data loading and preprocessing.

- Exploratory data analysis (EDA) with visualizations.

- Clustering for segmentation.

- Time series forecasting.

- Insights on sales trends, promotions, and store performance.

Tools and Technologies

- Python 3 (with libraries: Pandas, NumPy, Matplotlib, Seaborn, Scikit-learn, Statsmodels).

- Jupyter Notebook for development.

- Data source: CSV file ("Retail_Sales_Data_Unlox (1).csv").

Data Description

The dataset is loaded from a CSV file and contains transaction-level data. Key columns include:

- Date: Transaction date (e.g., 2023-04-13).

- Store_ID: Unique store identifier (e.g., STR_104).

- Store_Location: City (e.g., Chennai, Delhi).

- Product_ID: Unique product identifier (e.g., PRD_072).

- Product_Category: Category (e.g., Sports, Groceries, Home Appliances, Fashion).

- Product_Subcategory: Subcategory (e.g., Athletics, Outdoor, Household).

- Brand: Product brand (e.g., Reebok, Yonex, Nestle).

- Unit_Price: Price per unit.

- Units_Sold: Quantity sold.

- Total_Sales: Gross sales (Unit_Price * Units_Sold).

- Discount_Percentage: Applied discount (e.g., 5%, 15%).

- Revenue: Net revenue after discount.

- Customer_Type: New or Returning.

- Payment_Mode: Cash, UPI, Credit Card, Debit Card.

- Promotion_Applied: Yes/No.

- Stock_On_Hand: Remaining inventory.

- Store_Rating: Rating (e.g., 4.4).

- Region: North, South, East, West.

- Holiday_Flag: 0/1 indicator for holidays.

Methodology

1. Setup and Data Loading

- Google Drive is mounted for file access (in Colab environment).

- Libraries are imported for data manipulation (Pandas, NumPy), visualization (Matplotlib, Seaborn), clustering (KMeans, StandardScaler), and forecasting (SARIMAX).

- Data is loaded into a Pandas DataFrame and basic exploration (e.g., `df.head()`) is performed.

2. Exploratory Data Analysis (EDA)

- Visualizations likely include sales trends over time, category-wise revenue, store performance, etc. (Inferred from imports; specific plots not shown in the truncated notebook).

- Seasonal trends are analyzed using date-based grouping.

- Correlations between variables like discounts, promotions, and sales are explored.

3. Clustering (Segmentation)

- Features such as Total_Sales, Revenue, Store_Rating, Units_Sold, etc., are scaled using StandardScaler.

- KMeans clustering is applied to segment stores or products into groups (e.g., high-performing vs. low-performing stores).

- Elbow method or silhouette scores may be used to determine optimal clusters (not explicitly shown).

4. Time Series Forecasting

- Sales data is aggregated by date for time series analysis.

- SARIMAX model is fitted (accounting for seasonality, trends, and exogenous variables like holidays or promotions).

- Warnings are suppressed for clean output.

- Forecasts are generated for future periods to predict sales.

5. Visualization

- Plots use Matplotlib/Seaborn for trends, clusters, and forecast results (e.g., line plots for time series, scatter plots for clusters).

Results and Analysis

Key Findings from EDA

- Sales show seasonal peaks (e.g., higher during non-holiday periods based on Holiday_Flag).

- Top categories: Sports and Groceries contribute significantly to revenue.

- Discounts (5-15%) and promotions impact net revenue, with higher discounts leading to volume increases but margin reductions.

- Regional variations: South and North regions show diverse performance.

Clustering Results

- Stores are segmented into clusters based on performance metrics.

- High-performing clusters: Higher ratings (e.g., 4.4), better stock management, and higher revenue.

- Low-performing clusters: Lower ratings (e.g., 3.5), potential issues with inventory or pricing.

Forecasting Results

- SARIMAX model provides sales predictions, enabling demand planning.

- Example: Forecasted increases during peak seasons, with confidence intervals for uncertainty.

(Note: Specific numerical results and plots are not available in the truncated notebook; in a full execution, these would include generated figures.)

Business Insights & Recommendations

- **Seasonal Trends** : Sales peak in specific months; stock up inventory accordingly to avoid shortages.

- **Product Performance** : Certain categories (e.g., Sports, Groceries) drive revenue; prioritize marketing and stocking for these.

- **Store Segmentation** : Use high-performing stores (e.g., higher-rated ones in South region) as benchmarks for training/operations. For low-performers, optimize pricing, promotions, or inventory.

- **Promotions and Discounts** : Balance discounts to boost volume without eroding margins; analyze ROI on promotions.

- **Forecasting** : Use predictions for proactive planning, reducing overstock/understock risks.

- **Overall** : Implement AI-driven analytics for ongoing monitoring, supporting data-driven decisions in retail management.

Conclusion

This project demonstrates a comprehensive approach to retail analytics, from data loading to advanced modeling. By leveraging clustering and forecasting, retailers can gain a competitive edge through optimized operations and informed strategies. Future enhancements could include machine learning for demand prediction or integration with real-time data sources.

Appendix

- **Notebook File**: Retail_Analytics_and_Forecasting.ipynb

- **Data Source**: Retail_Sales_Data_Unlox (1).csv

- **References**: Python documentation for used libraries,Statsmodels for SARIMAX.