Retail Sales Forecasting & Customer Segmentation Analysis

1. Approach

The analysis focused on three primary tasks to derive actionable insights from the retail sales dataset:

Data Preprocessing

Missing data in both numerical and categorical columns was handled using median (for numerical) and mode (for categorical) imputation.

Exploratory Data Analysis (EDA) was performed to understand patterns, trends, and relationships across various features like sales, customer numbers, weather, promotions, etc. Relevant feature engineering was done, such as creating new features like log-transformed sales, scaled features, and binary weekend flags.

Data was encoded, scaled, and prepared for the predictive modeling process.

Sales Forecasting

Multiple machine learning models (Random Forest, XGBoost, SVM, Neural Networks) were used to predict daily sales (total_sales) for each store-category combination.

Hyperparameter tuning was conducted to optimize model performance using grid search with cross-validation.
The best performing model, Random Forest, was selected based on its R-squared of 0.9752 and Mean Squared Error (MSE) of 4462.22, showing its strong ability to forecast sales accurately.

The model was used to forecast sales for a 14-day period in January 2024.

Customer Segmentation:
Clustering (using K-Means) was applied to segment stores into meaningful groups based on features like total sales, number of customers, and transaction values.
Principal Component Analysis (PCA) was used to reduce dimensionality and visualize the segments.

The stores were grouped into distinct clusters with unique characteristics:

Cluster 1: Stores with higher in-store sales and lower online sales.

Cluster 2: Stores with a balanced sales distribution across both channels.

Cluster 3: Stores with strong online sales but lower foot traffic.

2. Findings

Sales Forecasting:

Random Forest outperformed other models, providing the best R-squared score, indicating it explains most of the variability in the sales data.
The model is able to provide reliable short-term sales forecasts, crucial for inventory planning, staffing, and promotions.

Customer Segmentation:
The segmentation revealed key differences in how stores operate (online vs. in-store sales, customer engagement, etc.).
These clusters offer actionable insights into how stores could tailor their strategies.
Stores with high in-store sales but lower online presence may benefit from improving their online offerings.
Stores that perform well online should focus on enhancing customer experience in physical stores to increase foot traffic.

3. Recommendations

For Sales Forecasting:
The Random Forest model can be used regularly for daily sales forecasting across stores to assist in decision-making.
Given the high accuracy, we recommend implementing this model into an automated pipeline for real-time sales prediction, which could assist in inventory and staffing decisions.

For Customer Segmentation:
Cluster 1: Focus on enhancing the in-store experience through better promotions, product placements, or events to drive more traffic.
Cluster 2: Implement balanced strategies to improve both online and in-store sales, focusing on cross-channel promotions.
Cluster 3: Increase efforts to improve the in-store experience, perhaps by introducing in-store exclusive offers or creating experiences that bring customers back to physical stores.

Further Steps:

Implement a feedback loop by continuously monitoring actual sales performance and adjusting model predictions accordingly.
Further deep dive into seasonality and other external factors like holidays to improve prediction accuracy, potentially using time series forecasting models.

4. Conclusion
This analysis provides valuable insights into how the retail chain can optimize both its sales forecasting and customer engagement strategies. With effective sales forecasting models and targeted segmentation strategies, stores can make data-driven decisions that increase efficiency and profitability.