Retail Sales Analysis

Introduction

This report describes the findings and methodologies used to address the retails sales analysis and machine learning using retail sales data from a multi-store retail chain. The challenge tasks require data preprocessing, sales forecasting and customer segmentation. This report will go into detail about the data preprocessing steps, model development and evaluation, findings and recommendations.

Problem Understanding

The challenge focuses on three primary tasks as defined:

a) *Data Preprocessing:* This task necessitates data cleaning such as handling missing values, performing exploratory data analysis to understand patterns and relationships, data preparation and feature engineering to improve predictive performance

b) *Sales Forecasting:* Develop a model to forecast the daily total sales for each store-category combination, create a 14-day forecast for January 2024, evaluate the model and analyze which features are most important for sales prediction

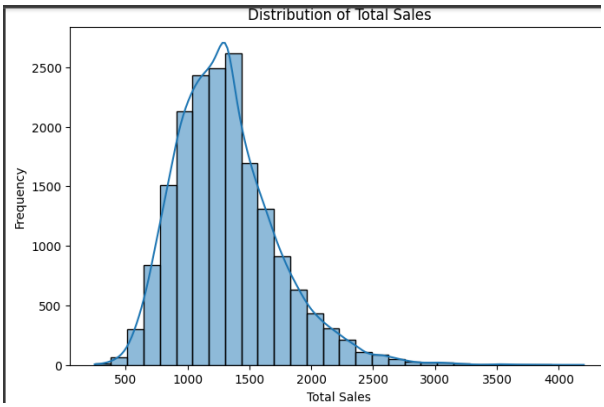
c) *Customer Segmentation:* Several metrics were used to assess the accuracy of the predictive models. In addition, alternative evaluation methods were considered and justified in terms of their relevance to the specific task requirements and dataset characteristics. Both coding implementations and this report were used to provide a thorough justification for the results.

Data Preprocessing

Following a thorough data exploration, several data preprocessing steps were made:

● *Null Values:* There were several null values in the data. The null values in the categorical variables were imputed using the variable mode while numerical variables were imputed using the variable mean.

● *Exploratory Data Analysis:* Summary statistics were observed showing the data distribution, seasonality and trend on the total sales were observed giving insight into the business. For instance, the total distribution of sales was observed as being positively skewed suggesting the possibility of one-time promotions or exceptional products. It could also mean very cheap products are underperforming or very expensive products are not being sold as much, as shown.



● *Feature Engineering:* To enhance the model's predictability, various feature engineering operations were employed. Temporal features like month and day captured seasonality, whereas lag features allowed the model to capture the trend of present sales. Rolling statistics reduced short-run volatility and highlighted underlying patterns. Event-based features allowed the model to factor in promotions or holidays that affected the sales pattern. The data was Finally, log-transforming total sales stabilized variance and improved the model's predictive performance across stores of different sizes of sales, improving accuracy and generalization.

● *Data Preparation*: Categorical variables were encoded to make them model-friendly, and numerical variables were scaled to achieve balanced influence among variables, to improve model performance.

Model Development

Sales Forecasting: Before proceeding with modeling, the features and target variable were defined. The dataset was separated into input features (X) and target variables (Y). The data was then divided into training and testing sets, with 80% for training (X_train, y_train) and 20% for testing (X_test and y_test). This splitting ratio ensures that there is enough data for model training while also allowing for the evaluation of the model's performance. Furthermore, the data splitting process ensures that the model is trained and evaluated on distinct subsets of data, preventing data leakage and providing an unbiased assessment of performance. A 14-day forecast was made on the prepared data and features analyzed to evaluate those that drive high sales. Average transactions was observed to be the most important feature that contributed the most to sales performance. The model used was Random Forest because of its ability to handle complex non-linear relationships, it is robust to overfitting and can easily identify key drivers of sales through ranking feature importance. It performed well with an accuracy of 97.87% and minimal MAPE error of 2.13%.

Customer Segmentation: For clustering, I classified store-level features like average daily sales, number of customers, transaction values, and online sales percentages along with return rates. After feature scaling to ensure equal contribution, I used the Elbow Method to identify the

optimal number of clusters, which showed 2 groups. KMeans clustering was
then used to divide stores based on performance as well as behavior patterns, allowing for tailored, store-specific strategies to be developed.

Key Insights

Customer Segmentation:

- Segment 0: The cluster is characterized with lower daily sales and a better online ratio.
- Segment 1: The cluster has stores with daily sales that are slightly higher. The return rates are also higher in comparison to cluster 0.

Sales Forecasting:

- Average transaction is a major contributor to the total sales.

Recommendations

1. For segment 0 there should be more investments on digital marketing to enhance online sales since they are relatively stronger.
2. For segment 1, local events such as community days should be increased to make use of the strong physical store and improving product descriptions to increase online guidance.
3. In the case of enhancing sales, stores should experiment with ways of ensuring there is an increase on customer expenditure per visit such as upselling companion products.