FUTURE INTERNS TASK-1 REPORT

(Sales Analysis & Predictive Modeling Dashboard)

Name: Taksh Samirkumar Patel

University: Pandit Deendayal Energy University

Course: Information and Communication Technology (ICT)

Semester: 4

Project Title: Sales Data Analysis & Machine Learning Dashboard

1. Introduction

The objective of Task-1 was to perform comprehensive exploratory data analysis (EDA) on a real-world sales dataset and to design an interactive dashboard that supports data-driven business decision-making. Additionally, a supervised machine learning model was implemented to predict future sales performance using historical transactional data, enabling stakeholders to forecast revenue and identify high-impact business drivers.

2. Dataset Overview

The dataset consists of detailed transactional sales records containing attributes such as sales amount, order date, profit, discount, quantity, product category and sub-category, geographic region, and customer-related fields. These attributes collectively enabled analysis of revenue distribution across time, products, and regions, as well as evaluation of profitability and customer behavior. The temporal fields supported the study of daily, monthly, and seasonal patterns in sales, while categorical fields (e.g., regions and product categories) allowed comparison of performance across different business segments.

3. Data Cleaning and Preprocessing

To ensure data quality, multiple preprocessing steps were applied before analysis and modeling. Missing values were identified and handled using appropriate strategies such as imputation or row removal, depending on the severity and business relevance of the records. Column names were standardized to a consistent, readable format to simplify downstream processing and avoid ambiguities during analysis. Date columns were converted to proper datetime format, and additional time-based features such as year, month, and day were derived to facilitate trend and seasonality analysis. Categorical variables (for example, product category and region) were encoded using suitable techniques to prepare the dataset for training the machine learning model. This systematic preprocessing pipeline ensured that the final dataset was clean, consistent, and suitable for both EDA and predictive modeling.

4. Exploratory Data Analysis (EDA)

EDA was conducted to understand the underlying structure and patterns present in the sales data. Revenue trends were examined at daily, monthly, and (where applicable) yearly levels to highlight growth patterns, peak sales periods, and low-performing intervals. Top-performing products, product categories, and regions were identified by aggregating metrics such as total sales and total profit, which helped in recognizing key revenue drivers and underperforming segments. Correlation analysis between numerical features (for example, sales, profit, discount, and quantity) was performed to understand how these variables interact and to detect potential multicollinearity issues relevant for modeling. Visualizations such as line charts, bar charts, and heatmaps were used to communicate insights clearly, making it easier for stakeholders to interpret trends and relationships.

5. Dashboard Development

An interactive sales analytics dashboard was developed using Streamlit to make insights accessible to non-technical users. The dashboard supports uploading input data in CSV and Excel formats, after which the data is automatically cleaned and processed using the predefined pipeline. Key performance indicators (KPIs) such as total revenue, total profit, average order value, and number of orders are displayed prominently to provide a quick overview of business performance. Revenue trends over time are visualized using time-series plots, allowing users to filter by date range, region, or product category. Categorical performance analysis is presented through bar charts and pie charts for dimensions such as region, category, and customer segment. A correlation heatmap is included to visually show relationships between numerical variables, and the dashboard also offers an option to download the cleaned dataset for further offline analysis.

6. Machine Learning Implementation

For predictive modeling, a Random Forest Regressor was employed to estimate future sales values based on historical data and engineered features. The dataset was split into training and testing subsets to objectively evaluate model performance and to avoid overfitting. Model accuracy and reliability were assessed using metrics such as R^2 Score, which measures the proportion of variance in sales explained by the model, and Mean Absolute Error (MAE), which quantifies the average magnitude of prediction errors. Feature importance analysis was carried out to identify the most influential predictors, such as product category, region, discount level, and temporal features. These insights help business users understand which factors most significantly impact sales outcomes and can guide strategic planning, promotional campaigns, and resource allocation.

7. Business Insights

From the combined EDA and modeling work, several key business insights were derived. Revenue was found to be heavily concentrated among a subset of top-performing products and categories, indicating opportunities for focused inventory and marketing strategies. Seasonal and time-based patterns were observed, showing that specific months or periods drive higher sales, which can inform demand forecasting and capacity planning. Certain regions consistently outperformed others in terms of both revenue and profit, suggesting potential for targeted regional strategies and expansion in high-growth areas. A strong positive correlation between profit and sales confirmed

that increases in sales volumes were generally aligned with higher profitability, subject to discounting and cost structures. The predictive model further demonstrated how data-driven forecasting can help businesses anticipate future sales and evaluate the impact of strategic decisions before implementation.

8. Conclusion

This project demonstrates practical proficiency in the complete data analysis pipeline, including data cleaning, exploratory data analysis, dashboard development, and machine learning-based prediction. The Streamlit dashboard provides an intuitive interface for monitoring sales KPIs, exploring trends, and downloading processed data, thereby bridging the gap between raw data and actionable insights. The Random Forest model shows how historical patterns can be leveraged to forecast future sales and identify key drivers of performance. Overall, the developed system equips organizations with a scalable, interactive tool to support continuous performance monitoring and informed, analytics-driven decision-making.