Project Report

1. Objectives

This project aimed to analyze retail sales data using advanced data science techniques and machine learning models. The primary goals were:

- To perform a **comprehensive Exploratory Data Analysis (EDA)** and understand key sales drivers, seasonal patterns, product demand, and customer behavior.
- To build **time series forecasting models** to predict future sales, aiding inventory management, demand planning, and business growth decisions.
- To develop an intelligent **product recommendation system** that suggests relevant products to users, increasing engagement and conversions.
- To compare different types of models—**statistical, machine learning, and deep learning**—for both forecasting and recommendation tasks, selecting the most effective solutions.

---

2. EDA Observations

A detailed EDA was conducted on the sales dataset containing over **110,000 rows** and **12 columns**. Key insights from the EDA are as follows:

1. Data Quality & Structure:

- Most variables were well-structured, including `CustomerID`, `ProductName`, `Price`, `Quantity`, and timestamp fields like `InvoiceDate`.
- **Missing values** were minimal and found mostly in customer and product-related columns. These were managed using filtering or imputation strategies.
- Some entries had **negative quantities**, indicating product returns or billing errors.

2. Distribution & Trends:

- `Quantity` was left-skewed — most purchases involved low volumes (1–5 units).
- `Price` showed a wide and uniform distribution, indicating products across various price ranges.
- **TotalSales** (calculated as `Quantity × Price`) had a right-skewed distribution with a few high-sale transactions.

3. Time-based Analysis:

- Extracting `Year` and `Month` from `InvoiceDate` revealed **monthly sales trends**.
- A **seasonal pattern** in sales was observed, with peaks in particular months — possibly due to discounts, festivals, or campaigns.
- Sales data was aggregated monthly for time series forecasting.

4. Product & Customer Behavior:

- Top-selling products accounted for a major share of the revenue.
- A small number of customers contributed significantly to total purchases, showing potential for **loyalty programs or personalized marketing**.

---

3. Approach (Modeling Pipeline)

The project followed a two-track modeling approach:

A. Time Series Forecasting

Target: Monthly total sales over time

Preprocessing Steps:

- Extracted `Year`, `Month` features from `InvoiceDate`
- Aggregated sales at the monthly level (`TotalSales`)
- Normalized and scaled values where needed

Models Used:

Model	Description
SARIMA	A classical statistical model capturing trend, seasonality, and autoregression.
Random Forest Regressor	Machine learning model using time-based engineered features (month, lag variables).
1D CNN (Convolutional Neural Network)	Deep learning model that captures temporal patterns through convolution over time-series sequences.

B. Recommendation System

Goal: Recommend relevant products based on customer purchase history and product associations.

Approaches Used:

Model	Description
KNN (Collaborative Filtering)	Based on user-product similarity using cosine distance from user-item matrix.
Autoencoder	A neural network that compresses customer behavior into a latent space and reconstructs likely purchases.
FP-Growth	Frequent pattern mining algorithm used for identifying co-purchased items (market basket analysis).

---

4. Model Comparison & Final Selection

Time Series Models Comparison:

Model	Performance	Remarks
SARIMA	Moderate accuracy, interpretable	Suitable for smooth seasonal data
Random Forest	Higher accuracy, captures non-linearities	Lacks temporal awareness
1D CNN	**Best performance**, lowest error (RMSE)	Captures sudden spikes, learns deep patterns

Selected Model: 1D CNN — due to its ability to learn complex patterns, adapt to fluctuations, and deliver strong predictive accuracy.

Recommendation Models Comparison:

Model	Strengths	Weaknesses
KNN	Easy to implement, intuitive	Poor performance in sparse data (cold-start problem)
Autoencoder	Learns hidden behavior, highly personalized	Needs tuning and training time
FP-Growth	Fast and interpretable product rules	Not personalized, rule-based only

Selected Models:

- **Autoencoder** (for personalized recommendations)
- **FP-Growth** (for co-purchase rule generation)

---

5. Key Findings

- There is a clear **seasonal sales trend**, likely driven by promotional events and customer buying behavior.
- **Loyal customers and top-selling products** significantly influence overall revenue.
- Advanced models like **1D CNN and Autoencoder** outperform classical models in accuracy and personalization.
- FP-Growth provides clear market insights into which products are frequently bought together.
- **Deep learning models** offered the best performance but required proper preprocessing and hyperparameter tuning.

---

6. Strengths, Weaknesses & Error Analysis

Strengths:

- Clean and diverse dataset with time, product, and user dimensions.
- Applied **hybrid modeling techniques** (statistical + ML + DL).
- Comprehensive feature engineering and visualization led to high-quality inputs for models.

Weaknesses:

- **Cold-start problem** in KNN where new users or products lack sufficient data.
- SARIMA limited by inability to handle non-linear trends or irregular fluctuations.
- CNN and Autoencoder models required GPU resources and training time.

Error Analysis:

- SARIMA underperformed during promotional peaks or unexpected surges.

- CNN sometimes struggled with extremely sharp spikes that lacked external context.
- KNN often gave **generic recommendations** for users with limited history.

---

7. Conclusion & Future Work

Conclusion:

The project successfully achieved its goals of forecasting sales and recommending products using a rich and diverse set of models. Insights from EDA allowed for accurate modeling, and the final selected models—**1D CNN** for forecasting and **Autoencoder + FP-Growth** for recommendations—demonstrated strong performance.

This system, when deployed, can help:

- Anticipate inventory demand
- Personalize customer experience
- Boost revenue through intelligent product placement

Future Enhancements:

- Incorporate **external variables** like promotions, holidays, and weather into time series forecasting for better accuracy.
- Use **RFM analysis** to segment customers and build targeted marketing strategies.
- Create a **hybrid recommender** that combines collaborative filtering with content-based filtering.
- Build a **real-time recommendation engine** integrated into an e-commerce dashboard (using Streamlit, Flask, or Power BI).
- Automate periodic **model retraining** using MLOps tools.
