AI-Driven Demand Forecasting & Replenishment — Project Report

I. Introduction

1.1 Problem Statement

Modern retail and supply-chain operations struggle with accurately forecasting product demand, especially in environments influenced by seasonal trends, promotions, and unpredictable market behavior. Traditional spreadsheet-based or rule-based forecasting methods often fail to capture these complexities, resulting in:

- **Stockouts**, which lead to lost sales and decreased customer satisfaction
- **Overstocking**, which ties up working capital and increases holding costs
- **Inefficient manual planning**, requiring significant time and human intervention
- **Reactive inventory decisions**, rather than proactive and data-driven strategies

These challenges highlight the need for a more intelligent, automated system that can continuously learn from historical patterns and adjust to changing demand dynamics.

1.2 Project Goal

The goal of this project is to design and implement an **AI-driven demand forecasting and automated replenishment solution** capable of delivering accurate SKU-level predictions and actionable inventory insights. The system aims to:

- Utilize an **LSTM-based forecasting model** to predict short-term demand
- Provide **automated reorder recommendations** using ROP (Reorder Point) and Safety Stock formulas
- Offer a **live dashboard** for visualization of forecasts, trends, and risk alerts
- Expose a **REST API** for seamless integration with ERP or procurement systems
- Enable faster, more informed inventory decisions with reduced manual effort

This project serves as a proof of concept demonstrating how advanced analytics and machine learning can modernize traditional supply-chain workflows.

1.3 Digital Transformation Context

This solution directly supports the broader shift toward **Industry 4.0 and digital supply-chain transformation**. Organizations worldwide are transitioning from manual, intuition-driven decision making to automated, data-driven operational models. The project aligns with these trends by:

- Introducing AI for **predictive forecasting**, reducing uncertainty in inventory decisions
- Replacing spreadsheet-based workflows with **real-time dashboards** and intelligent alerts
- Supporting **API-driven integrations**, enabling system scalability across business units
- Increasing supply-chain agility through continuous learning and faster response to demand changes

By adopting such an AI-enabled framework, businesses can significantly improve forecasting accuracy, optimize inventory levels, and enhance overall operational efficiency.

II. Solution Architecture

2.1 Architecture:

Data → ETL → Feature Store → LSTM Model → API → Dashboard → Automated Reorder Engine

2.2 Technology Stack:

Layer	Technologies	Justification
Data Ingestion & ETL	Python, Pandas, NumPy	Mature ecosystem for data cleaning, excellent for time-series preprocessing.
Feature Engineering	Scikit-learn, custom transforms	Provides scaling, splitting, and robust ML utilities.
Forecasting Model	TensorFlow/Keras LSTM	Capable of learning temporal dependencies, outperforming classical ARIMA for multi-feature time-series.
API Layer	FastAPI or Flask	Lightweight, async-friendly, ideal for ML inference serving.
Dashboard	Streamlit	Quick interactive visualization; easily deployable.
Cloud & Deployment	Render	Containerization ensures reproducibility; cloud enables scalability.

2.3 Technology Stack Justification:

- **LSTM chosen over ARIMA/ETS** because the dataset contains promotions, seasonality, and nonlinear patterns which LSTM handles better.
- **FastAPI** provides high-performance inference and easy documentation via OpenAPI.
- **Streamlit dashboard** accelerates development and allows business stakeholders to interact with forecasts immediately.
- **Cloud deployment** reduces infrastructure maintenance and enables elastic scaling for inference-heavy workloads.

III. Proof of Concept (PoC) & Implementation

3.1 Dataset Description

The PoC uses a simplified retail dataset representing:

- Daily sales per SKU
- Inventory levels per SKU
- SKU metadata (lead times, cost, categories)
- Calendar signals (weekday, holiday)

Common issues addressed:

- Missing dates
- Zero-demand periods
- Seasonality and variance across SKUs
- Promotion-driven spikes

3.2 ETL & Feature Engineering

Key transformations:

- Date normalization and merging calendar features
- Rolling averages (7-day, 30-day)
- Lag features (1–14)
- Standard scaling
- Per-SKU (or per-volume bucket) normalization for stable model convergence

3.3 ML Model Implementation

The model is a **Sequence-to-Multi-Step LSTM**:

- Input window: 30 days
- Prediction horizon: 7–14 days
- Architecture: LSTM → Dense layers → Horizon output
- Loss: MSE, R2 Score, MAE
- Optimizer: Adam
- Model enhancements attempted:
    - Early Stopping
    - Validation split using time-aware partitioning

    - Evaluation metrics (MAE, RMSE, MAPE)

3.4 Deployment Environment

- Model saved as a TensorFlow artifact.
- API endpoint /predict deployed via FastAPI + Docker.
- Dashboard visualizes:
    - Forecast curves
    - Actual vs predicted
    - Reorder alerts
    - Inventory risk heatmap

3.5 PoC Functionality

The PoC successfully:

- Loads historical SKU data
- Preprocesses it into sequences
- Trains LSTM and outputs forecasts
- Calculates ROP + Safety Stock
- Shows these results on a dashboard

3.6 PoC Limitations

- Limited historical data reduces generalization.
- No full walk-forward cross-validation in the PoC phase.
- Uncertainty estimation (P10/P50/P90) only partially implemented.
- Sparse SKUs not treated with specialized intermittent models.
- No real connection to ERP for automated POs — simulated only.

IV. Business Model & Value

4.1. Digital Model:

- Transforms manual forecasting into predictive, semi-autonomous inventory management.

4.2. Automated Inventory Optimization

- Maintain optimal stock levels, reducing manual effort and minimizing errors.

4.3. Significant Cost Savings

- Minimize holding costs, prevent stock-outs, and reduce waste through intelligent reordering.

4.4. Enhanced Operational Efficiency

- Streamline inventory processes, freeing up resources for strategic initiatives.

4.5. Scalable & Adaptable

- Easily scale the solution to manage inventory across multiple products and store locations.

4.6. Data-Driven Decision Making

- Empower retail managers with accurate insights for better purchasing and sales strategies.

4.7. Time Savings from Automation

- Automate repetitive tasks, allowing teams to focus on value-added activities.

V. Risk & Governance

5.1. Cybersecurity Measures

- **HTTPS Enforcement:** All data transmitted between the dashboard, API, and backend services is encrypted using HTTPS to prevent interception.
- **Secure Deployment:** The solution is deployed using secure cloud configurations, including private networking, firewall rules, and restricted inbound connections.

5.2. Data Privacy & Compliance

- **SKU-Level Data Only:** No customer-level or personally identifiable information (PII) is used in the model, reducing regulatory exposure.
- **Encrypted Storage:** All stored data, model artifacts, and logs are encrypted at rest using cloud-native encryption mechanisms.
- **Compliance Alignment:** The architecture follows standard organizational data-handling practices and aligns with general privacy best practices (GDPR-safe due to no personal data collection).

5.3. Ethical & Operational Risks

- **Model Drift Monitoring:** Forecast accuracy is monitored over time to detect when the model begins to degrade due to new trends or shifts in demand patterns.
- **Human Oversight:** The system includes mechanisms for managers to review approve, or override AI-generated reorder recommendations.
- **Transparency:** Decision logs explain why a reorder alert was generated (e.g., dropping below ROP), ensuring clarity for users.
- **Versioned Governance:** Both datasets and model versions are recorded, allowing full traceability of model updates and decision logic changes.

5.4. Auditability & Governance Framework

- **Model Version Control:** Every model update is documented and versioned to ensure reproducibility and accountability.
- **Action Logging:** Key operations—such as forecast generation, reorder calculations, and manual overrides—are logged for auditing.
- **Governance Practices:** The system maintains documentation of ETL processes, forecasting methodology, and decision rules, supporting internal governance and future audits.