AI-Powered Demand Forecasting & Inventory Optimization Platform

1. Business Goal

The goal is to provide a SaaS platform that enables retailers and wholesalers to optimize demand forecasting and inventory management. By leveraging advanced machine learning models, the platform helps businesses reduce stockouts, minimize overstock, improve cash flow, and increase customer satisfaction.

2. System Architecture

The system is designed as a modular microservice-based architecture with the following components:

- Data Ingestion Layer (CSV, APIs, DB connectors)
- Data Storage (Cloud storage like S3/GCS + PostgreSQL/TimescaleDB)
- ETL Pipelines (Airflow/Prefect + dbt)
- Feature Store (Feast or PostgreSQL)
- ML Training Service (Prophet + LightGBM + PyTorch models)
- Model Registry & Tracking (MLflow)
- Model Serving Layer (FastAPI + Docker + Kubernetes)
- Monitoring & Drift Detection (EvidentlyAI, Prometheus, Grafana)
- Frontend Dashboard (React + Tailwind + Recharts)

3. Data Flow & Storage

Raw sales data flows into a staging layer (CSV/DB). ETL pipelines clean and transform the data before moving it into the feature store. Training datasets are versioned and stored in cloud storage. Predictions are stored in a time-series database for analysis and dashboard visualization.

4. ML Pipeline

- Training Pipeline: Data validation → Feature extraction → Model training (Prophet, LightGBM) → Hyperparameter tuning (Optuna) → Register in MLflow
- Prediction Pipeline: Load latest model → Batch/real-time inference → Post-process with business rules → Store results → Expose via API/dashboard

5. Deployment & Infrastructure

- CI/CD with GitHub Actions (build Docker → push registry → deploy via Helm to K8s)
- Dev → Staging → Prod environments
- Horizontal autoscaling (HPA), DB partitioning for multi-tenant support

6. Monitoring & Maintenance

- Model drift detection with EvidentlyAI
- System metrics tracked via Prometheus + Grafana
- Forecast accuracy monitored through backtesting
- Error alerts via Slack/Email integrations

7. Security & Compliance

- TLS encryption for data in transit, KMS for data at rest
- Role-based access control (JWT/Auth0)
- GDPR-compliant retention policies
- Audit logging for governance

8. Tech Stack

- Data: PostgreSQL/TimescaleDB, S3/GCS, dbt
- ML: Prophet, LightGBM, PyTorch, Optuna, MLflow
- Serving: FastAPI, Celery, Redis
- Infra: Docker, Kubernetes, Terraform, AWS/GCP/Azure
- Observability: Prometheus, Grafana, ELK stack
- Frontend: React, TailwindCSS, Recharts

9. Future Enhancements

- Upgrade to advanced deep learning models (Temporal Fusion Transformers, DeepAR)
- Support for real-time streaming forecasts (Kafka + Spark)
- Multi-cloud deployment capability
- Online learning for continuous model updates