Pharma Sales Forecasting Project Plan

Project Overview

This project aims to develop a robust, data-driven sales forecasting pipeline tailored for the pharmaceutical industry. The goal is to predict weekly or monthly sales values at the brand and product levels while incorporating the effects of promotions, pricing, and regional variations. The outcome will include a fully documented pipeline and an interactive Streamlit dashboard for scenario analysis.

Dataset Overview

- Source: Pharma Sales Data (Kaggle)
- Records: Approximately 600,000 (2014–2019)
- Granularity: Brand / Product / Customer / Date
- Key Variables: Date, Product_ID, Customer_ID, Country, Sales_Value, Sales_Quantity, Price_per_Unit, Discount, Region, Channel, Brand, Therapeutic_Area
- Target Variable: Sales_Value (TRY or USD)
- Frequency: Weekly (resampled from daily if needed)

Technical Stack and Tools

- Data Cleaning: pandas, numpy, pyjanitor
- Visualization: matplotlib, seaborn, plotly
- Feature Engineering: scikit-learn, tsfresh, feature-engine
- Modeling: statsmodels (SARIMAX), prophet, scikit-learn, xgboost
- Hierarchical Reconciliation: scikit-hts or custom MinT implementation
- Evaluation Metrics: WAPE, sMAPE, RMSE
- Dashboard: Streamlit, Plotly
- Documentation: Markdown, Sphinx, or mkdocs
- Reproducibility: pytest, makefile, requirements.txt

7-Day Project Plan

Day 1 — Data Understanding and Repository Setup

- Create project structure and import dataset.
- Inspect columns, data types, and missing values.
- Draft data dictionary and README with project context.
- Define target variable (Sales_Value) and forecast horizon (8 weeks).
- Initialize src/config.py and src/data_prep.py scripts.
- Deliverables: data_dictionary.md, README.md draft, clean dataset, requirements.txt.

Day 2 — Data Cleaning and Aggregation

- Convert date column to datetime format and resample to weekly frequency.

- Handle missing and inconsistent values.
- Aggregate by Brand, Channel, Region, and Country.
- Validate pricing consistency (Value = Quantity × Price).
- Implement data quality tests for nulls, duplicates, and outliers.
- Deliverables: processed dataset, data_quality_report.md.

Day 3 — Exploratory Data Analysis and Feature Engineering

- Analyze trends, seasonality, and price elasticity.
- Visualize brand-level heatmaps and promo vs sales correlations.
- Define feature set: lags, rolling stats, calendar features, promo/discount flags.
- Document feature plan in docs/feature_plan.md.
- Deliverables: 01_eda.ipynb, feature_plan.md, src/features.py.

Day 4 — Baseline Models (Naive, sNaive, SARIMA, SARIMAX)

- Perform train/test split (last 3 months for testing).
- Implement Naive, Seasonal Naive, SARIMA, and SARIMAX models.
- Use rolling-origin cross-validation (5 folds × 8-week horizon).
- Evaluate performance using WAPE, sMAPE, and RMSE.
- Deliverables: 03_models_baseline.ipynb, baseline_results.md.

Day 5 — Machine Learning Models (XGBoost, LightGBM, CatBoost)

- Generate lag and rolling window features.
- Transform data into supervised learning format.
- Train and tune models using GridSearchCV or Optuna.
- Analyze feature importance with SHAP or permutation analysis.
- Compare ML results with baseline models.
- Deliverables: 04_models_ml.ipynb, xgb_model.pkl, SHAP plots.

Day 6 — Hierarchical Forecasting and Streamlit Dashboard

- Construct hierarchical structure: SKU → Brand → Region → Country.
- Implement Bottom-Up and MinT reconciliation methods.
- Develop Streamlit dashboard with price and promo sliders for scenario analysis.
- Display KPIs (WAPE, RMSE) and allow CSV export.
- Deliverables: Streamlit app, forecast figures, updated README.

Day 7 — Evaluation, Documentation, and Packaging

- Perform backtesting and summarize performance results.
- Document modeling assumptions, limitations, and risks in modeling_card.md.
- Finalize README with business interpretation and reproduction guide.

- Create Makefile and requirements.txt for reproducibility.
- Prepare GitHub repository for portfolio presentation.
- Deliverables: final README, modeling_card.md, fully reproducible project.

Business Summary

This project applies both classical time series models and modern machine learning techniques to forecast pharmaceutical sales at multiple hierarchical levels (SKU, Brand, Region, Country). It demonstrates analytical and business acumen through structured documentation, reproducibility, and actionable insights. The final deliverables provide a comprehensive, realistic example of how data science can support demand planning, supply optimization, and brand strategy in the pharmaceutical industry.