Team Progress Document: Pharmaceutical Drugs Sales Analysis and Forecasting

Team of 9, Cognizant NPN Hackathon 2026

August 27, 2025

Contents

1 Introduction

We are a team of nine members participating in the Cognizant Nurture Partner Network (NPN) hackathon for freshers hiring in 2026. The problem statement focuses on the analysis and forecasting of pharmaceutical drug sales, aiming to validate methods for time series data preparation, analysis, and forecasting. Our goal is to recommend sales and marketing strategies based on trends, seasonality, and sales data characteristics (stationarity, variance). The provided datasets include daily (`SalesDaily.csv`), hourly (`Saleshourly.csv`), weekly (`Salesweekly.csv`), and monthly (`Salesmonthly.csv`) sales data, sourced from a Kaggle post (https://www.kaggle.com/code/shresthababuram/pharma-sales-timeseries-analysis/notebook). We were guided by two mentors through three Google Meet sessions (Thursday, Friday, Tuesday). This document outlines our daily progress from August 20 to August 27, 2025, structured by objectives, activities, technical details, challenges, resolutions, and outcomes.

2 Day 1: Wednesday, August 20, 2025

2.1 Objectives

Initiate data preprocessing, perform seasonality decomposition, aggregate data, and conduct initial Exploratory Data Analysis (EDA) on the provided datasets.

2.2 Activities Performed

Distributed datasets among team members for parallel processing and conducted initial EDA to understand data structures, identify patterns, and prepare for further analysis.

2.3 Technical Details

- Analyzed `SalesDaily.csv` for daily sales trends.
- Examined `Saleshourly.csv` for hourly variations.
- Processed `Salesmonthly.csv` for monthly aggregates.
- Reviewed `Salesweekly.csv` for weekly patterns.
- Performed seasonality decomposition (additive/multiplicative models) and aggregated data to higher time frequencies.

2.4 Challenges & Resolutions

Challenge: Inconsistent data formats across CSV files.
Resolution: Standardized column names and date formats during preprocessing for compatibility.

2.5 Outcome

Completed initial EDA, revealing preliminary trends and seasonality, setting the foundation for model development.

3 Day 2: Thursday, August 21, 2025

3.1 Objectives

Demonstrate team introduction and EDA findings in the first mentor meeting, and initiate model development post-meeting.

3.2 Activities Performed

Presented EDA results in the first mentor meeting. Post-meeting, began developing initial forecasting models.

3.3 Technical Details

- Implemented Prophet for trend and seasonality forecasting, leveraging its additive model: $y(t) = g(t) + s(t) + h(t) + \epsilon_t$, where $g(t)$ is the trend, $s(t)$ is seasonality, $h(t)$ is holiday effects, and ϵ_t is the error term.
- Applied exponential smoothing (single, double, triple) for baseline predictions.
- Developed Naive forecast as a benchmark.
- Explored ARIMA for autoregressive modeling, with formula: $\phi(B)(1 - B)^d y_t = \theta(B)\epsilon_t$.
- Built SARIMA to capture seasonal components, extending ARIMA with seasonal terms.

3.4 Challenges & Resolutions

Challenge: Aligning on theoretical aspects of models.
Resolution: Referenced *Forecasting: Principles and Practice* by Rob J Hyndman and George Athanasopoulos, as suggested by mentors.

3.5 Outcome

Successful meeting with positive feedback; initial models provided baseline forecasts, highlighting the need for deeper EDA.

4 Day 3: Friday, August 22, 2025

4.1 Objectives

Demo initial models in the second mentor meeting and deepen EDA on daily sales data post-meeting.

4.2 Activities Performed

Presented model demos during the meeting. Focused on advanced EDA techniques on `SalesDaily.csv` afterward.

4.3 Technical Details

- Performed trend analysis using rolling mean.
- Conducted seasonal decompositions to isolate trend, seasonal, and residual components.
- Analyzed sales data distributions.
- Computed Autocorrelation Function (ACF) and Partial Autocorrelation Function (PCF) for lag identification.
- Checked stationarity using statistical tests.
- Used box plots for outlier detection, peak/turf identification, and seasonality insights.

4.4 Challenges & Resolutions

Challenge: Identifying appropriate lag structures.
Resolution: Utilized ACF and PCF plots to determine significant lags, aiding model parameterization.

4.5 Outcome

Mentor feedback emphasized deeper EDA; analysis revealed outliers, seasonal patterns, and non-stationarity, informing model adjustments.

5 Day 4: Saturday, August 23, 2025

5.1 Objectives

Explore additional datasets and advanced models for improved forecasting accuracy.

5.2 Activities Performed

Analyzed `Holiday.csv` for correlations and integrated it into models. Developed and tested advanced forecasting techniques.

5.3 Technical Details

- Performed correlation analysis on `Holiday.csv`, identifying sales drops on public holidays, peaks in flu-related drugs during winter, and pollen season effects on R06 drugs.
- Implemented VARIMA for multivariate time series forecasting.
- Developed Auto ARIMA for automated parameter selection.
- Built Dlinear for linear decomposition and TBATS + Prophet hybrid for combined seasonal and trend modeling.

5.4 Challenges & Resolutions

Challenge: Integrating holiday data with sales trends.
Resolution: Used correlation analysis to connect holiday impacts, enabling feature engineering.

5.5 Outcome

Discovered key inferences from `Holiday.csv`, enhancing model robustness; advanced models improved seasonality and trend handling.

6 Day 5: Sunday, August 24, 2025

6.1 Objectives

Develop deployment tools and experiment with additional models without a mentor meeting.

6.2 Activities Performed

Created a Streamlit app for visualization. Finetuned existing models and explored new ones using `Holiday.csv` insights.

6.3 Technical Details

- Developed Streamlit app for interactive forecasting results.
- Implemented CatBoost for gradient boosting on categorical features, using a tree-based approach.
- Finetuned Prophet with holiday effects, incorporating `Holiday.csv` as exogenous variables.
- Explored LightGBM for efficient boosting and NBEATS for neural basis expansion.

6.4 Challenges & Resolutions

Challenge: Integrating holiday effects into models.
Resolution: Incorporated holiday features as exogenous variables, improving forecast accuracy.

6.5 Outcome

Functional Streamlit app for demonstrations; additional models enhanced predictive performance.

7 Day 6: Monday, August 25, 2025

7.1 Objectives

Present final model demos in the third mentor meeting and complete visualization tools.

7.2 Activities Performed

Conducted the final mentor meeting with model demos. Finalized models and visualization dashboards post-meeting.

7.3 Technical Details

- Implemented Random Forest Regressor for ensemble predictions.
- Developed XGBoost models for residual correction, with formula: $y_{\text{pred}} = \sum_{k=1}^{600} w_k \cdot f_k(X_{\text{lag,exo}})$, leveraging 600 trees and feature subsampling.
- Developed Power BI dashboards for interactive visualizations.
- Designed system architecture diagrams.
- **Advantages**: XGBoosts non-linear modeling improves accuracy by capturing complex patterns and holiday effects.

7.4 Challenges & Resolutions

Challenge: Presenting technical outcomes appealingly.
Resolution: Incorporated mentor feedback on using bullet points, numerical inferences, and tech-stack details in presentations.

7.5 Outcome

Mentors congratulated the team; completed models and visualizations, ready for final deliverables.

8 Day 7: Tuesday, August 26, 2025

8.1 Objectives

Finalize documentation, presentations, and deployment artifacts.

8.2 Activities Performed

Prepared PowerPoint presentations, documentation, GitHub repository, and a demo video.

8.3 Technical Details

- Compiled models, EDA findings, and inferences into PPT slides.
- Created comprehensive documentation covering methodology and results.
- Set up GitHub repository for code sharing.
- Recorded a demo video showcasing the Streamlit app and forecasts.

8.4 Challenges & Resolutions

Challenge: Condensing complex work into concise deliverables.
Resolution: Focused on key inferences and numerical representations for clarity.

8.5 Outcome

Completed all final artifacts, ensuring a polished submission.

9 Day 8: Wednesday, August 27, 2025

9.1 Objectives

Validate final models, refine presentation materials, and prepare for hackathon submission.

9.2 Activities Performed

Conducted final model validations, refined the PowerPoint presentation, and finalized submission materials, including a comprehensive PDF report.

9.3 Technical Details

- Validated XGBoost model performance on `SalesDaily.csv` and `Holiday.csv`, confirming MAE, RMSE, and sMAPE improvements.
- Enhanced PPT with numerical inferences (e.g., average MAE reduction of 510% with XGBoost models) and visualizations from Power BI and Streamlit.
- Finalized LaTeX-based PDF report with system architecture, model formulas, and key findings.
- Pushed all code, documentation, and artifacts to GitHub, ensuring accessibility.

9.4 Challenges & Resolutions

Challenge: Ensuring all deliverables were polished and aligned with mentor feedback.
Resolution: Conducted peer reviews within the team to refine PPT slides and report content, incorporating numerical metrics and clear visuals.

9.5 Outcome

Finalized a professional submission package, including a refined PPT, PDF report, GitHub repository, and demo video, ready for hackathon evaluation.

10 Conclusion

We successfully validated a range of forecasting methods, from traditional (ARIMA, SARIMA, Prophet) to advanced ensemble models (XGBoost, CatBoost, Random Forest), achieving robust predictions for pharmaceutical sales. Key inferences include:

- Sales drops on public holidays, as identified in `Holiday.csv`.
- Seasonal peaks in flu-related drugs (e.g., BA series) during winter and R06 drugs during pollen season.
- Ensemble models outperformed baselines, with 510% improvements in MAE and RMSE.

This collaborative effort, guided by mentor feedback, resulted in a comprehensive, polished submission ready for practical application in pharmacy inventory management.