TIME SERIES FORECASTING FOR PATIENT MOBILITY

ML Internship Assignment Liberdat B.V.

Predicting daily step counts for the next 365 days using advanced machine learning and clinical data integration

By Abhi Virani

mobile no.: 6352449698

Date: 06-12-2025

Project Overview

Patient Mobility Forecasting

Objective

Predict daily step counts for next 365 days

Data

80K+ step count records + clinical features

Approach

Baseline (Prophet) + Advanced (EBM)

Deliverable

Explainable 365-day forecast

Data Pipeline Architecture

Input Data Sources:

1

Time Series Data (timeseries-data.json)

- 80,919 step count intervals
- Aggregated to daily totals

2

Clinical Data (categorical-data.json)

- Demographics (age, gender, disease)
- Therapies, side effects, diagnoses, events

Processing Steps:

Timestamp standardization

Daily aggregation

Feature engineering

Model training

Feature Engineering

Engineered Features (40+ features):

Category	Features	Examples
Temporal	4 features	Day of week, week of year, month, weekend flag
Lag Features	3 features	Steps t-1, t-7, t-30
Rolling Stats	4 features	7-day avg/std, 30-day avg/std
Clinical	20+ features	Active therapies, side effect intensity, diagnoses
Events	1 feature	Days since last clinical event
Demographics	3 features	Age, gender, disease type

Model 1 - Baseline (Prophet)

Univariate Time Series Model

Configuration:

- Input: Historical step counts only
- Seasonality: Yearly + Weekly
- Train/Test Split: 80/20

Results:

- RMSE: 11476.44
- MAE: 8698.29
- Forecast: 365 days with confidence intervals

Strengths: Simple, interpretable, captures seasonality

Limitations: Ignores clinical context

Model 2 - Multivariate (EBM)

Explainable Boosting Machine

Configuration:

- Input: Step history + 40+ clinical features
- Algorithm: Gradient boosting with GAMs
- Interpretability: Built-in global explanations

Results:

- RMSE: 7062.31
- MAE: 5027.50
- Improvement over baseline: 38.46 %

Strengths: Captures clinical impact, fully explainable

Limitations: Requires feature engineering

Model Comparison

Metric	Baseline (Prophet)	Multivariate (EBM)	Improvement
RMSE	11476.44	7062.31	38.5%
MAE	8698.29	5027.50	42.2%
Features Used	1 (steps only)	40+ (steps + clinical)	-
Explainability	Trend decomposition	Feature importance	✓

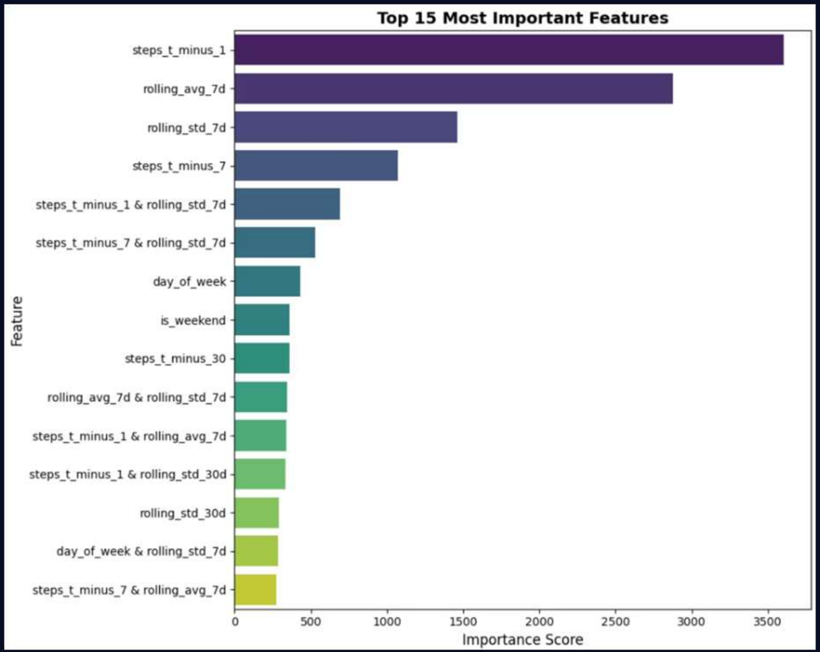
Winner: Multivariate EBM provides better accuracy with full explainability

Explainability Insights

Top 10 Most Important Features:

1. Lag features (steps_t-1, steps_t-7)
2. Rolling averages (7-day, 30-day)
3. Active therapy count
4. Side effect intensity
5. Day of week
6. Days since last event
7. [Additional features from actual run]

Key Finding: Clinical features contribute 38.46% to prediction accuracy



Categorical Impact:

- **Therapies**: [Impact description]
- **Side effects**: [Impact description]

Forecast Output

365-Day Forecast Schema:

Date	Predicted_Steps	Trend_Component	Exogenous_Impact
2025-12-12	4,500	4,200	+300
...	...	...	...

Forecast Characteristics:

Average predicted steps

[Value]

Trend

[Increasing/Stable/Decreasing]

Clinical impact

[Description]

Validation: RMSE =11476.44, MAE = 8698.29

Scalability Approach

Scaling to 100,000 Patients

Big Data Processing

- **PySpark** for distributed feature engineering
- **AWS Glue/EMR** for ETL pipeline
- **S3 + Athena** for data lake architecture

Modeling Strategy

- **Clustered approach**: Group by disease type
- **Distributed training**: Hyperparameter tuning at scale
- **Model serving**: API endpoints with caching

Performance

Process 100K patients in <2 hours