Activity Forecasting — Daily Steps Prediction

Using Time-Series & Clinical Features

Author: *Sakshi Gharat — ML Intern*

Tools: Python · Google Colab · Prophet · EBM (InterpretML)
Deliverables: 365-day forecast · Explainability · Clean Pipeline
Date: 12/12/2025

Notebook Link: https://colab.research.google.com/drive/1NHHyNk1WoWJQ9BEhI-OHC2CkHjlmfYSi?usp=sharing

Problem Statement & Objective

Problem Statement

1. Wearable devices generate high-frequency activity data, but raw step counts are **noisy** and **incomplete**.
2. Clinical context (age, smoking status, therapy) is often ignored in **traditional** time-series forecasting.
3. Healthcare applications require **accurate forecasts with explainability,** not black-box predictions.

Project Objective

- Forecast **daily step counts for the next 365 days** for a single patient.
- Build:
    1. **Baseline univariate time-series model** using historical steps.
    2. **Multivariate explainable ML model** incorporating clinical features.
- Ensure:
    - **Clean**, reproducible data pipeline
    - Quantitative evaluation (**RMSE, MAE**)
    - Transparent and interpretable predictions suitable for healthcare use.

Data Pipeline Overview

Data Sources

- **Time-Series Data (File A)**Wearable step-count events with timestamps (`start`, `end`, `count`)
- **Clinical Data (File B)**Patient-level attributes (age, smoking status, therapies, diagnoses)

Preprocessing Steps

- Converted all timestamps to **Python datetime objects**
- Standardized timezone to **UTC**
- Removed duplicate and malformed records
- Sorted events chronologically

Daily Aggregation

- Aggregated raw step events into **Daily_Step_Count**
- Ensured **continuous daily index** (filled missing days with zero activity)
- Produced a clean, gap-free daily time series

Pipeline Output

- One row per day
- Aligned temporal and clinical data
- Ready for feature engineering and modeling

Feature Engineering & Clinical Fusion

Time-Series Features

- **Lag Features**
  - `steps_t-1` : Previous day step count
- Captures short-term temporal dependency in activity patterns

Calendar Features

- Daily aggregation aligned on calendar date
- Enables downstream seasonal and trend modeling

Clinical Feature Fusion

- **Age:** Derived from birth year
- **Smoking Status:** Binary indicator (`is_smoker`)
- **Therapy Exposure:** Binary indicator (`is_on_therapy`)

Design Decision

- Clinical records lacked reliable event-level timestamps
- Therefore, **static snapshot-based fusion** was used
- Prevents incorrect temporal assumptions or data leakage

> *This approach prioritizes data integrity over artificial feature inflation.*

Pipeline Output

Each day contains:

- `daily_steps`
- `steps_t-1`
- `age`, `is_smoker`, `is_on_therapy`

Baseline Model: Univariate Time-Series Forecasting

Model Choice: ***Prophet***

Model Inputs

- ds → Date
- y → Daily step count

> No clinical or exogenous features used in this model

Training Strategy

- Model trained on historical daily step counts
- Future dataframe generated for **365-day horizon**

Validation Setup

- **Hold-out validation window** used on recent history
- Metrics computed on unseen test data

Evaluation Metrics

- **RMSE** — penalizes large forecast errors
- **MAE** — average absolute deviation

(Used as baseline comparison for Model 2)

Outputs

- **Daily forecasts for next 365 days**
- Decomposition into **trends, weekly / yearly seasonality, overall prediction (**yhat**)**

Baseline Model Results & Performance

Performance Metrics(Prophet)

Metric	Value
RMSE	~11,476
MAE	~8,698

Observed Behavior

- Captures long-term **trend,** weekly and yearly seasonality
- Struggles with:
  - Sudden activity drops or spikes
  - Health-related behavior changes (not modeled here)

Forecast Characteristics

- Smooth predictions driven by historical patterns
- No awareness of **therapy status, smoking behavior, clinical events**

Limitations of Baseline

- Assumes activity depends **only on past steps**
- Cannot model patient-specific health context
- Motivates the need for **multivariate modeling**

Multivariate Model: Feature Set & Training

Model Choice: ***Explainable Boosting Machine (EBM)***

Input Feature Set

Time-Series Features

- `steps_t-1` (previous day steps)
- Calendar alignment via daily indexing

Clinical Features (Static)

- `age`
- `is_smoker`
- `is_on_therapy`

Target Variable: Daily Step Count

Training Methodology

- Dataset split using **temporal ordering**
  - Early data → training
  - Recent data → validation
- Model trained using squared loss objective

Model Performance Comparison

Evaluation Setup

- **Validation Strategy:**
  - Temporal hold-out split
  - Last ~20% of days used as test set
- **Metrics Used:**
  - **RMSE (Root Mean Squared Error)**
  - **MAE (Mean Absolute Error)**

Performance Metrics

Model	RMSE	MAE
Model 1: Prophet (Univariate)	~11,476	~8,698
Model 2: EBM (Multivariate)	~7,646	~5,544

Observations

- EBM significantly reduces **error magnitude**
- Incorporating **clinical features improves accuracy**
- Multivariate model captures variability not explained by trend alone

Model Explainability (EBM Insights)

Why Explainability Matters

- Healthcare forecasting requires **transparent and interpretable models**
- Stakeholders must understand **why** predictions change, not just the output

Explainability Method

- **Model:** Explainable Boosting Machine (EBM)
- **Technique Used:**
  - Global feature importance
  - Additive feature contribution analysis (interpretml)

Key Feature Impacts (Global)

- **Lag Feature (steps_t-1):**
  - Strongest positive predictor
  - Recent activity strongly influences next-day steps
- **Age:**
  - Higher age → gradual decrease in predicted activity
- **Therapy Status:**
  - Active therapy associated with reduced step counts
- **Smoking Status:**
  - Minor but consistent negative impact

Interpretation

- EBM provides **human-readable, monotonic effects**
- Clinical variables modify baseline activity trends
- Model behavior aligns with **medical and behavioral intuition**

Takeaway

> EBM enables both **accurate prediction** and **transparent clinical interpretation**, making it suitable for healthcare applications.

Final Output & Deliverables

Model Comparison Summary

Model	Type	RMSE	MAE
Prophet	Univariate Baseline	Reported	Reported
EBM	Multivariate	Improved vs Baseline	Improved vs Baseline

Forecast Output

- **Horizon:** 365 days
- **Granularity:** Daily
- **Output format (CSV):**
  - Date
  - Predicted_Steps
  - Trend_Component
  - Exogenous_Impact

Deliverables Submitted

- ✅ Google Colab notebook (end-to-end pipeline)
- ✅ 365-day forecast CSV
- ✅ Supporting methodology document
- ✅ Scalability & cloud-readiness document
- ✅ Presentation deck (this file)

Conclusion:

This solution demonstrates a production-ready, interpretable time-series forecasting pipeline tailored for healthcare activity data