Supporting Document

1. Key Challenges in the Data Engineering Pipeline

The first challenge was aggregating high-frequency step data into a clean daily format. The second challenge involved handling ongoing clinical events with missing end dates. The third challenge was resolving timezone inconsistencies across multiple data sources.

2. Modeling Approach and Justification

A SARIMA model was used as a baseline to capture trend and seasonality in step counts. An Explainable Boosting Machine (EBM) was then used as the primary model because it provides strong performance while remaining interpretable. The model used lagged step features, therapy indicators, side-effect severity, and calendar-based features. Performance was evaluated using RMSE and MAE metrics.

3. Learnings from the Explainability Phase

Explainability analysis showed that past step counts were the strongest predictors of future activity. Therapy periods generally increased mobility, while side effects and clinical events reduced step counts. These insights increased confidence in the model and validated real-world clinical intuition.

Conclusion

This project demonstrates an end-to-end, interpretable machine learning pipeline for patient mobility forecasting, suitable for real-world healthcare applications.