**LSTM (Long Short-Term Memory)**

* **What it is:** A type of recurrent neural network (RNN) that is designed to learn patterns in sequential data.
* **Why we use it:** Member health data often comes as time-based sequences (e.g., daily or weekly activity, hospital visits, lab results). LSTM can remember long-term trends and detect patterns that may indicate risk of health deterioration.
* **How it works in this project:**
  + **Takes sequences of member activity over 30, 60, and 90-day windows.**
  + **Learns temporal patterns, such as declining engagement, repeated hospital visits, or missed appointments.**
  + **Produces a temporal embedding, which is a condensed representation of each member’s recent health activity.**
* **Output:** A vector of features that summarizes the member’s sequential health behavior, ready to be used by the next model (XGBoost).

**Key benefit:** Captures short-, medium-, and long-term trends in member behavior that static models can’t see.

**XGBoost (Extreme Gradient Boosting)**

* **What it is:** A type of decision tree-based ensemble model that combines multiple weak predictors to form a strong prediction.
* **Why we use it:**
  + XGBoost is excellent at handling tabular data with both numeric and categorical features.
  + It can integrate the temporal embeddings from LSTM with static member features (like age, membership type, or region).
  + Produces a risk score for each member, which can then be used for tier-based stratification.
* **How it works in this project:**
  + **Receives input: LSTM embeddings + static features.**
  + **Outputs: Risk score for 30/60/90-day windows.**
  + **Risk scores are used to classify members into five risk tiers: Very High, High, Moderate, Low, Minimal.**

**Key benefit:** Combines dynamic temporal patterns and static features to generate highly accurate risk predictions for stratification.

**Why Combine LSTM and XGBoost?**

1. **LSTM captures temporal trends in sequences.**
2. **XGBoost excels at predictive classification using structured data.**
3. **Together, they create a hybrid model that is:**
   * **Accurate over multiple time horizons (30/60/90 days).**
   * **Able to produce actionable risk tiers.**
   * **Prototype-ready and scalable for production.**

**Recommended Model Setup**

**1. LSTM (Sequence Model)**

* **Purpose: Learn temporal patterns from member activity over different time windows.**
* **Approach:**
  + **Use three separate LSTM models, one for each window:**
    1. **30-day LSTM → captures short-term risk patterns.**
    2. **60-day LSTM → captures medium-term trends.**
    3. **90-day LSTM → captures long-term deterioration patterns.**
* **Why separate: Each window has distinct temporal dynamics. Training separate models improves prediction accuracy for each horizon.**

**2. XGBoost (Classifier)**

* **Purpose: Combine embeddings from each LSTM with static member features (age, membership type, region) to predict risk tier.**
* **Approach:**
  + **Use three separate XGBoost models, one for each window:**
    1. **Embeddings from 30-day LSTM → XGBoost #1**
    2. **Embeddings from 60-day LSTM → XGBoost #2**
    3. **Embeddings from 90-day LSTM → XGBoost #3**
* **Why separate: Allows stratification specific to each prediction horizon, which is actionable for care teams.**

**Model Training Phase: Detailed Explanation**

**The model training phase is where the predictive system learns to identify members at risk of health deterioration. It consists of two main steps: pretraining and fine-tuning.**

**1. Pretraining on Synthetic Data**

* **Purpose: Teach the model general patterns of member health deterioration before using real data.**
* **Why synthetic data: Real member data may be limited or incomplete. Synthetic data allows the model to learn trends safely without privacy concerns.**
* **What it contains: Simulated member activity over time, demographics, and risk outcomes.**
* **How it works:**
  1. **The model analyzes sequences of member activity (e.g., transactions, visits, lab results) over 30, 60, and 90-day periods.**
  2. **It learns temporal patterns—for example, declining engagement or increased hospital visits—that often precede health deterioration.**
* **Result: A model that has basic knowledge of how members’ health can change over time.**

**2. Fine-Tuning on Real Data**

* **Purpose: Adapt the pretrained model to actual member behaviors to improve prediction accuracy.**
* **How it works:**
  1. **The model takes the real member data, structured in the same 30/60/90-day windows.**
  2. **It adjusts its internal rules based on real trends, such as specific demographic risks, seasonal patterns, or local healthcare practices.**
  3. **The model generates temporal embeddings, which are condensed representations of each member’s recent health activity.**
* **Integration with static features: The embeddings are combined with static member features (age, membership type, region) for richer input.**
* **Outcome: A refined model capable of predicting short-term, medium-term, and long-term deterioration with high accuracy.**

**3. Window-Based Training**

* **30-day window: Captures immediate risk signals, identifying members who may deteriorate quickly.**
* **60-day window: Captures medium-term trends, balancing short-term fluctuations and longer trends.**
* **90-day window: Detects slow deterioration patterns over time.**
* **Why this matters: Training separate windows ensures the model can predict deterioration across different time horizons, making risk stratification actionable for care teams.**

**4. Model Output**

* **After training, the model can:**
  + **Generate a risk score for each member over 30, 60, or 90 days.**
  + **Feed these scores into the XGBoost component (or similar classifier) to stratify members into five risk tiers: Very High, High, Moderate, Low, Minimal.**
* **Key benefits:**
  + **Accurate early detection of high-risk members.**
  + **Flexible prediction horizons for proactive care planning.**
  + **Foundation for actionable, tier-based interventions.**

**Member Risk File**

* A separate **CSV/JSON file** is generated containing:
  + Member ID
  + 30/60/90-day risk scores
  + Assigned risk tier
  + Optional suggested interventions
* Purpose: Provides a **self-contained dataset** for dashboard visualization, testing, and simulation.