Forecasting ILI Trends Using Machine Learning

Author: Hieu Pham
Advisor: Dr. Bo Mei

TCU College of Science & Engineering RESEARCH
Science and Engineering Research Center

OBJECTIVE

This project aims to build an interactive forecasting tool for influenza-like illness (ILI) using historical data and machine learning models. The app allows public health analysts and researchers to visualize trends and make predictions for better outbreak preparedness.

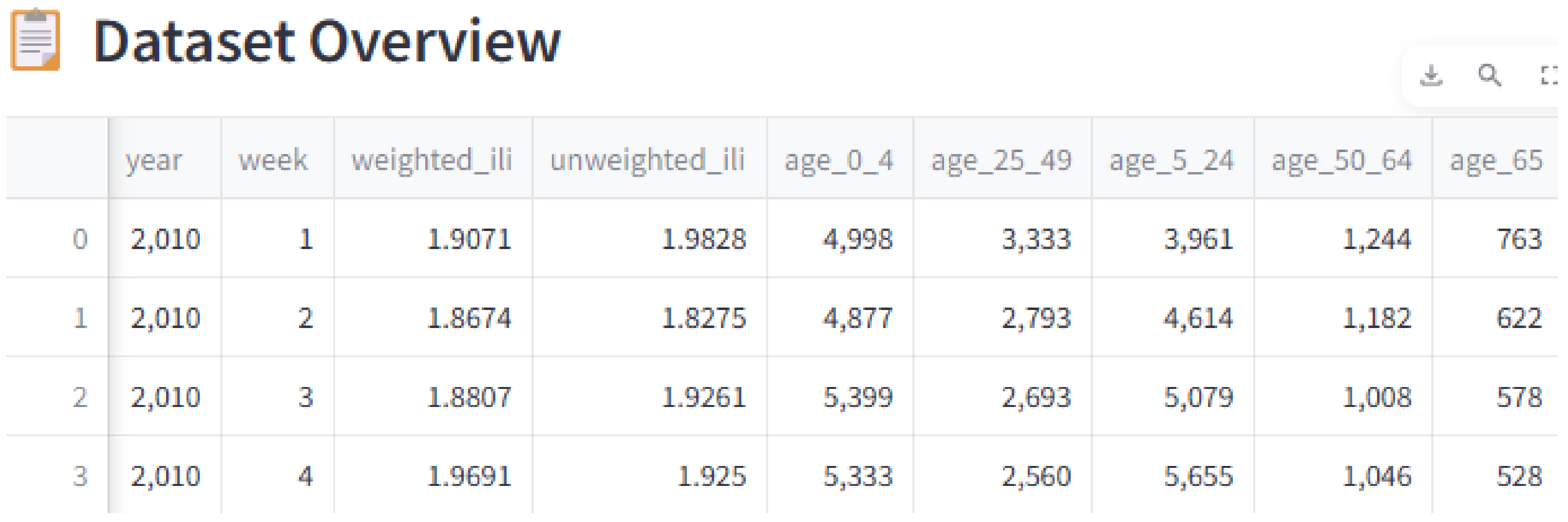
DATASET OVERVIEW

Data is directly downloaded from CDC Flu View, which includes weekly ILI data from 2010 onwards in the United States on a national level.

Some crucial characteristics include:

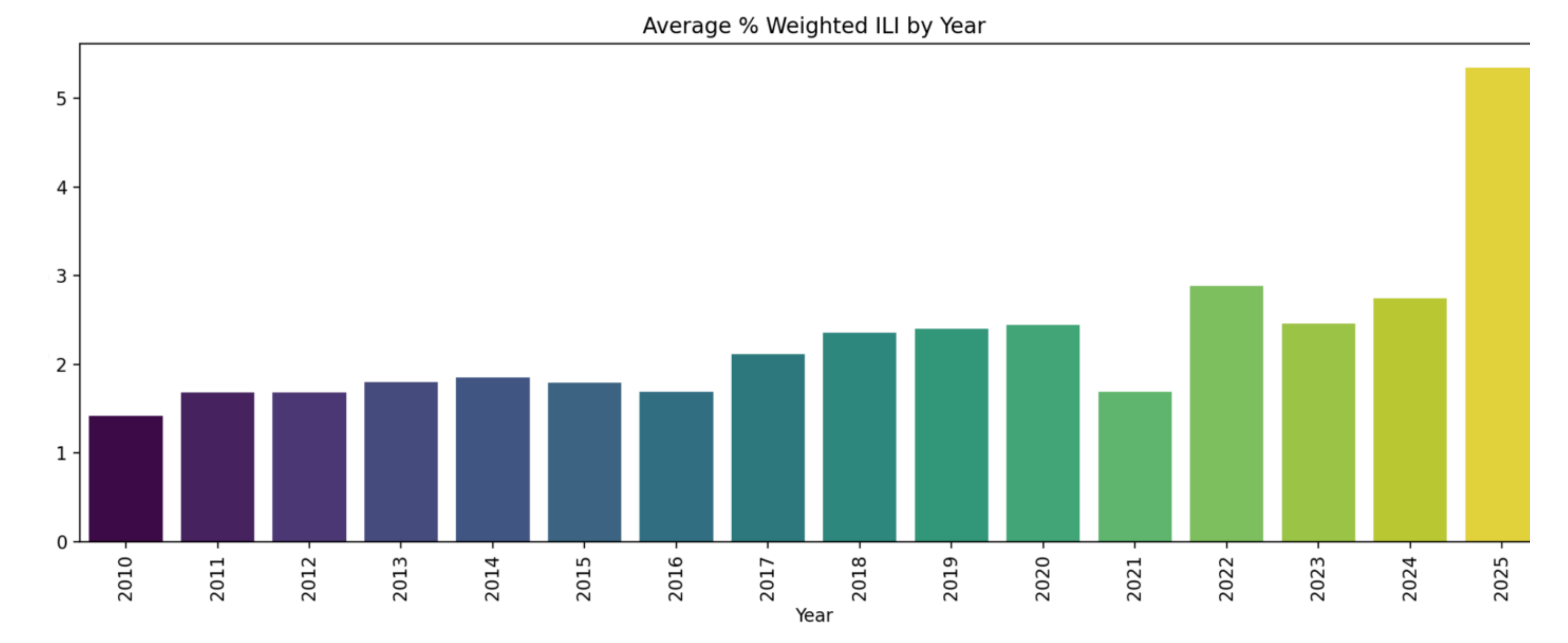
- Date of report (Year, week)
- Weighted % ILI cases (Used for ML models)
- Age-group breakdown
- Total patient counts

We also created lag features for autoregression (ILI_t-1 to ILI_t-5), date-based aggregations (year, week) to support the machine learning process.



Dataset Overview

	year	week	weighted_ili	unweighted_ili	age_0_4	age_25_49	age_5_24	age_50_64	age_65
0	2,010	1	1.9071	1.9828	4,998	3,333	3,961	1,244	763
1	2,010	2	1.8674	1.8275	4,877	2,793	4,614	1,182	622
2	2,010	3	1.8807	1.9261	5,399	2,693	5,079	1,008	578
3	2,010	4	1.9691	1.925	5,333	2,560	5,655	1,046	528



ML MODELS USED

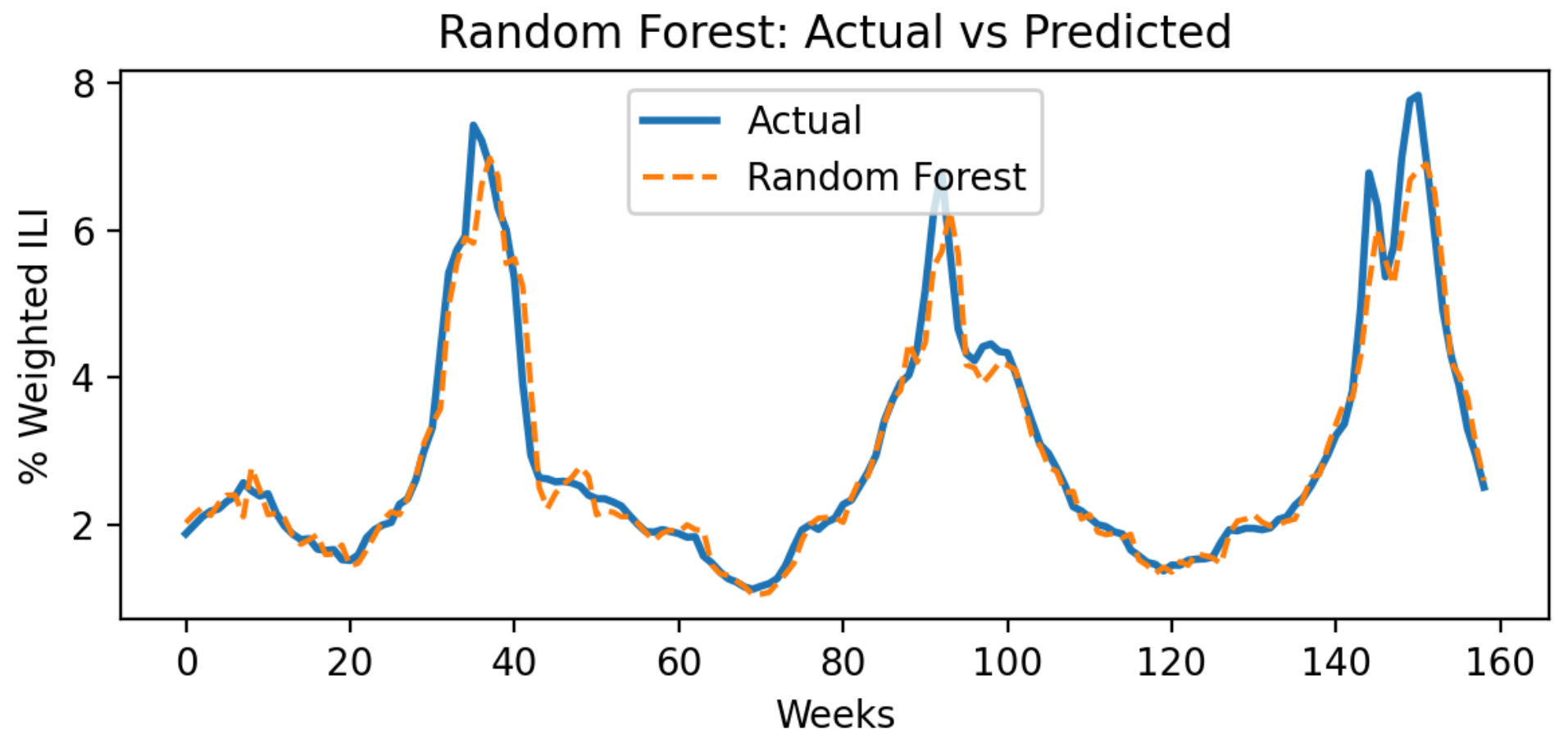
- ***Random Forest***
  - Handles non-linear data, robust to overfitting
  - Slower training, less interpretable
- ***Linear Regression***
  - Simple, fast, great for linear trends
  - Struggles with complex patterns
- ***XGBoost***
  - High accuracy, handles missing data
  - Complex and slower to train

EVALUATION METRICS

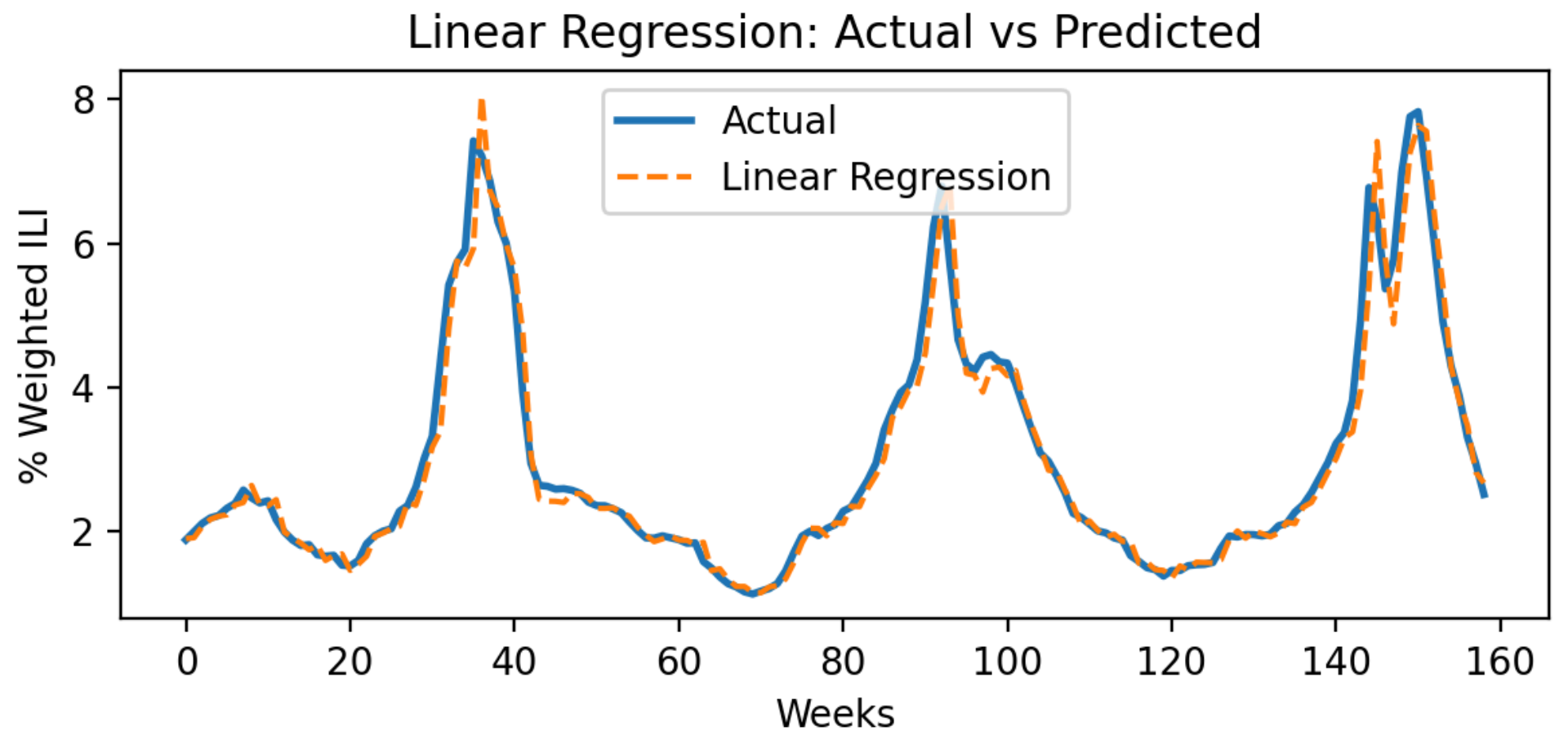
- *RMSE (Root Mean Squared Error)*
  - To know how accurate the predictions are.
  - Example: RMSE = 0.50 → predictions are off by 0.5% on average.
  - ⇒ **Lower is better**
- *Use R^2 Score (Coefficient of Determination)*
  - To know how well the model fits the data.
  - $R^2 = 1.0$ → *perfect prediction*
  - ⇒ **Higher is better**

MODEL EVALUATIONS

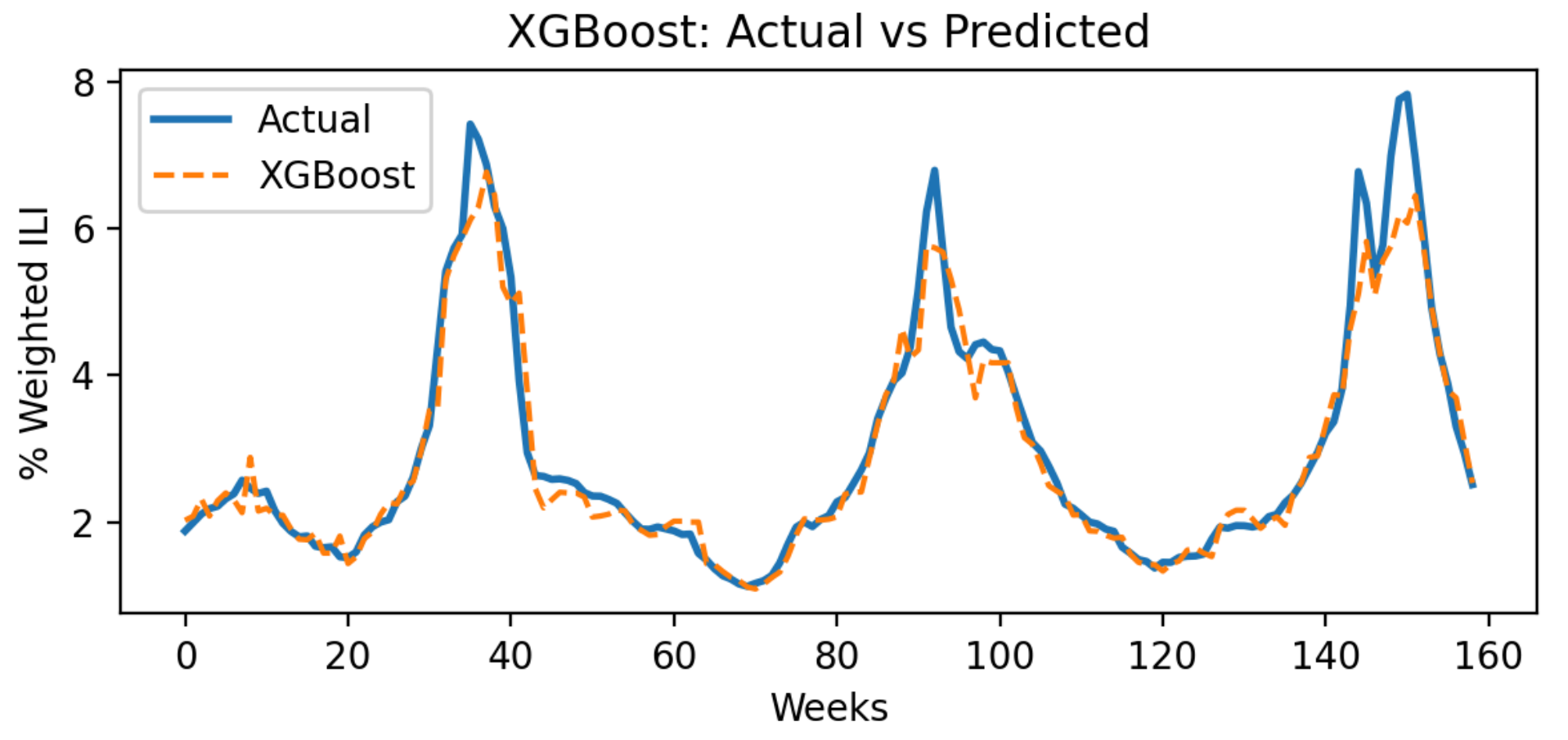
- ***Random Forest***
  - RMSE: 0.128
  - R^2 Score: 0.951



- ***Linear Regression***
  - RMSE: 0.106
  - R^2 Score: 0.960

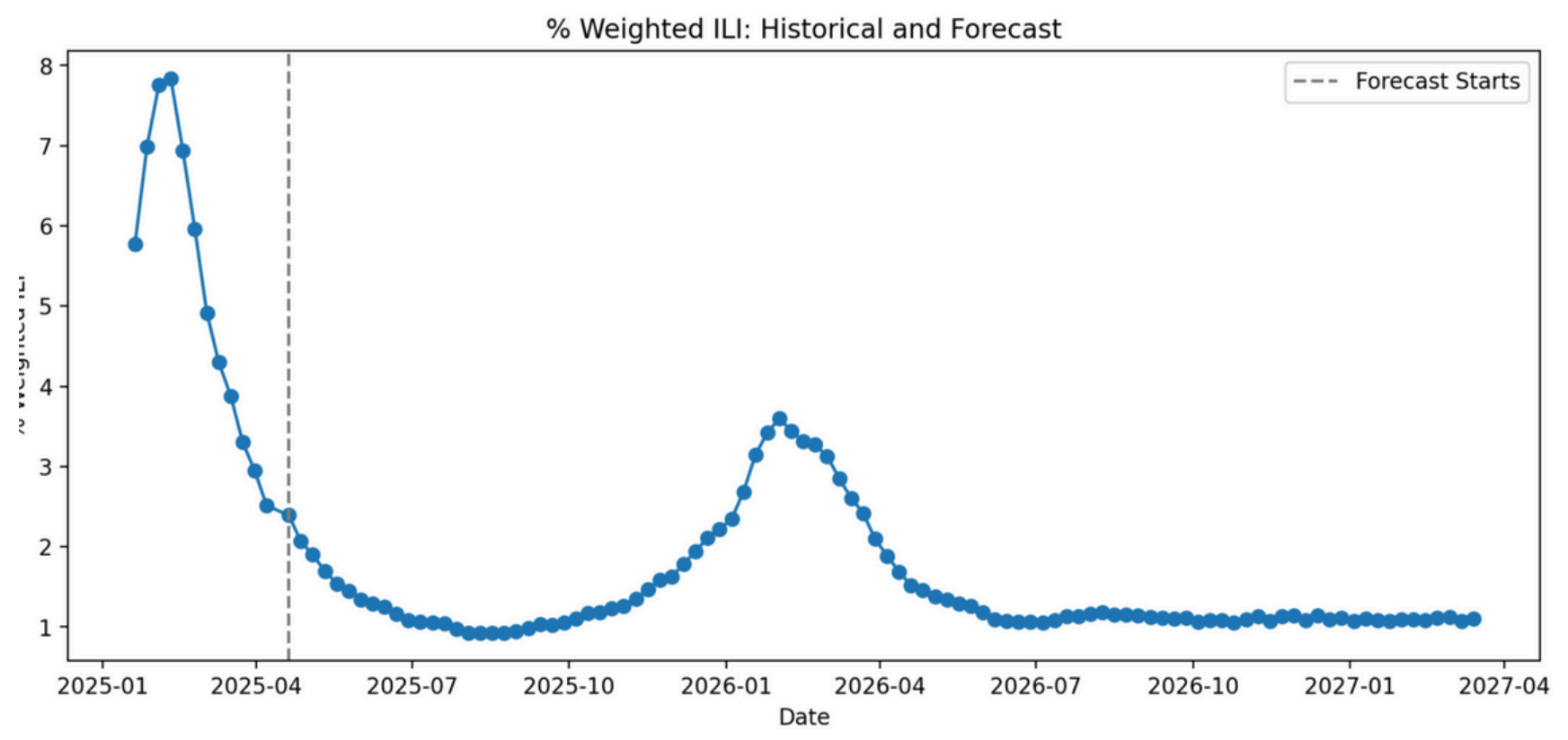


- ***XGBoost***
  - RMSE: 0.148
  - R^2 Score: 0.944

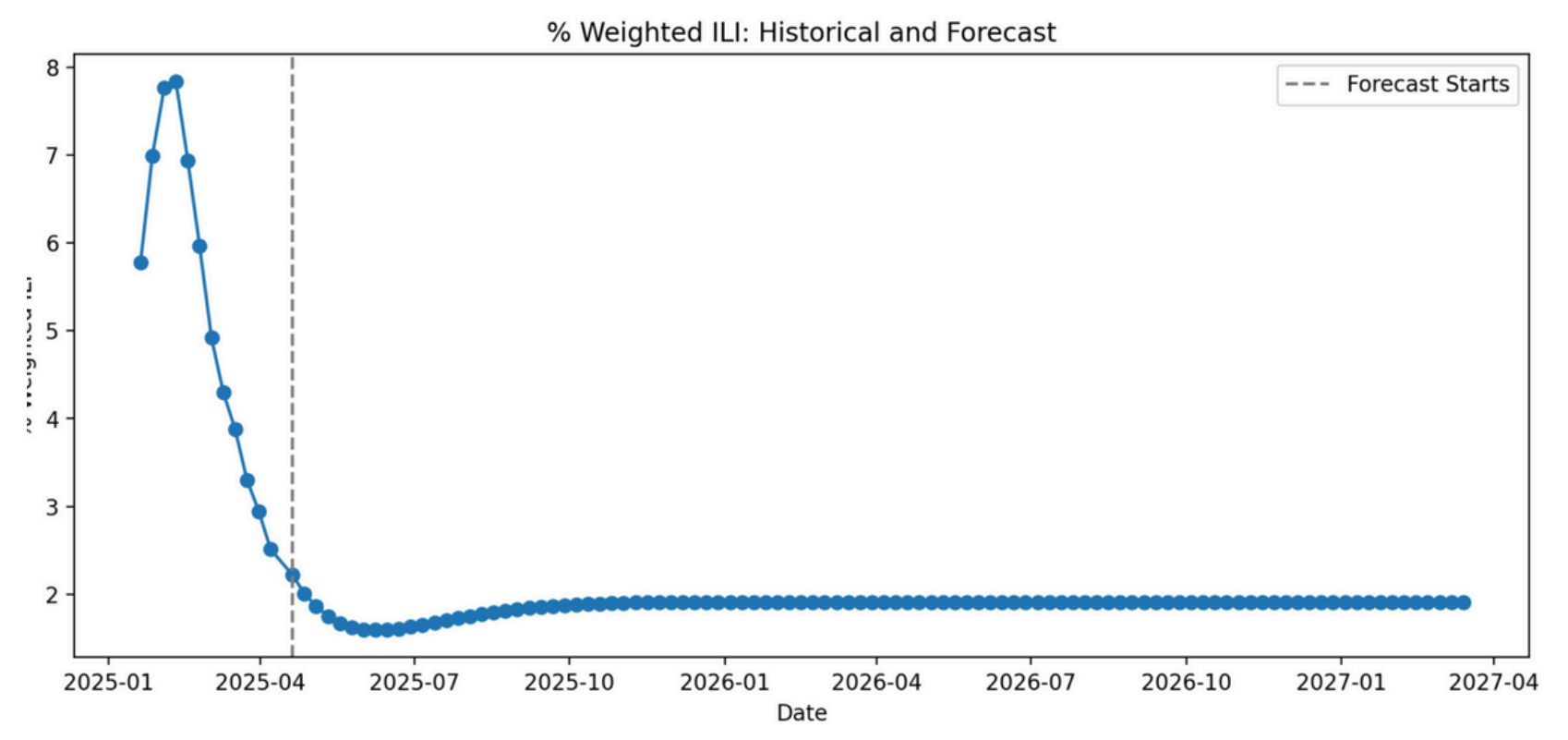


KEY TAKEAWAYS AND FORECAST

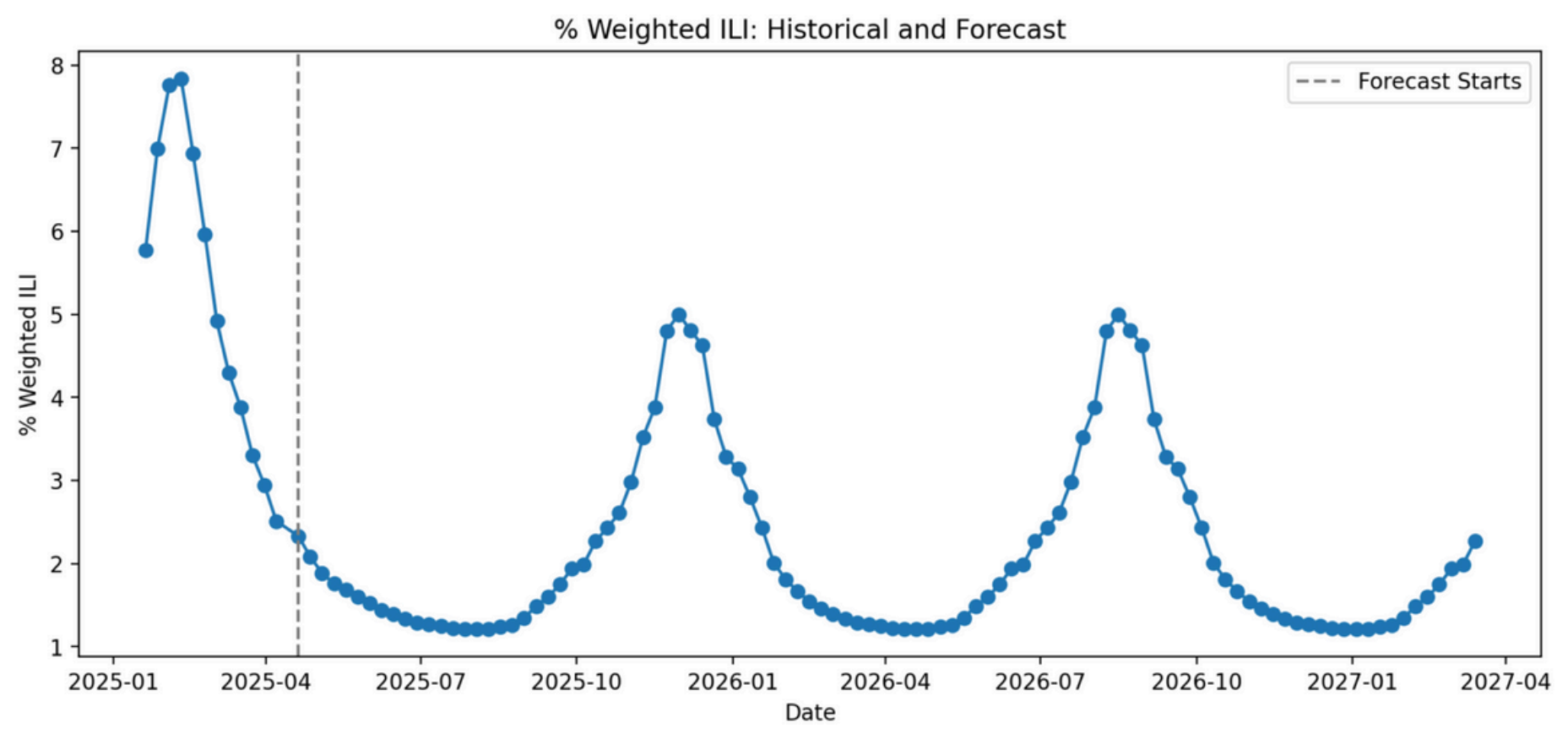
Linear Regression performed best in both RMSE and R². Despite being a simple model, it effectively captured the trend in the ILI data — likely because the target has a mostly linear structure.



Random Forest



Linear Regression



XGBoost

For more detailed and interactive plots, check out the deployed app for the project!

TOOLS AND TECH

pandas

Python & Pandas

Streamlit

Streamlit (Web app)

matplotlib

Matplotlib & Seaborn

scikit learn

Scikit-learn, XGBoost

ACKNOWLEDGEMENTS

I sincerely appreciate Dr. Bo Mei for the valuable feedback throughout the design process, helping me stay focused on realistic key goals, and thoughtfully addressing my questions.