A Time-Series Approach to Predicting CVE Volume

Jerry Gamblin

Presented at the vuln4cast colloquium

Darwin College, Cambridge

CVEforecast.org

I am pleased to announce the launch of a rogolabs project focused on predictive CVE volume forecasting. This innovative platform is now operational, and I extend a warm invitation for you to explore its functionalities. I particularly encourage you to formulate questions and insights, which we can discuss at the end of this session.

The Scaling Crisis

CVE volume isn't just growing—it's exploding. Security professionals face an unprecedented challenge managing vulnerability disclosure rates that show no signs of slowing.

Traditional vulnerability management remains fundamentally reactive, forcing teams into perpetual firefighting mode rather than strategic defense planning.

From Reactive to Predictive

1

Traditional Approach

React to known vulnerabilities
Fix what's already exposed
Always behind the curve

2

Forecasting Approach

Predict future disclosure patterns
Allocate resources proactively
Stay ahead of the threat landscape

Time-Series Analysis: The Foundation

Definition

Time-series analysis is a statistical methodology employed to analyze time-ordered sequences of data points. Each observation is indexed by a temporal parameter, enabling the identification of underlying dynamics, dependencies, and structural properties inherent in the chronological progression of a variable. This method is crucial for understanding stochastic processes, forecasting future values, and performing inferential analysis on temporal phenomena across various domains such as econometrics, meteorology, and signal processing.

Core Components

- **Trend:** The big picture – the consistent upward or downward movement of data over a long period. Think of a company's steady growth or a gradual decline in temperatures over decades.
- **Seasonality:** Predictable, repeating patterns that occur at regular intervals. Like holiday sales spikes, daily traffic jams, or annual weather cycles.
- **Noise:** The unpredictable, erratic ups and downs in data that don't fit into trends or seasonal patterns. These are the random quirks and unexpected events.

CVEforecast: Project Philosophy

CVEforecast transcends traditional tooling—it's a self-improving, automated platform engineered for production environments. The core challenge lies in achieving accuracy and reliability within cybersecurity's dynamic landscape.

Built on open-source principles, it embodies the collaborative spirit of academic research while delivering enterprise-grade performance.

The Data Pipeline Challenge

Massive Scale Operations

Managing 300,000+ CVE JSON files daily requires sophisticated data engineering. The pipeline handles:

- Real-time data ingestion
- Automated preprocessing
- Quality validation checks

`code/data_loader.py` orchestrates this complex workflow, ensuring data integrity throughout the forecasting pipeline.

The Ensemble Modeling Strategy

Statistical Models

Prophet, ARIMA, LinearRegression, Croston, ExponentialSmoothing

Machine Learning

XGBoost, LightGBM, RandomForest

Deep Learning

TCN, NHiTS, Theta, TiDE

15+ models working in concert deliver superior accuracy compared to any single approach.

The Self-Tuning Brain

The system continuously evolves through `code/tuner/comprehensive_tuner.py`—a sophisticated optimization engine that never stops learning.

Performance Tracking

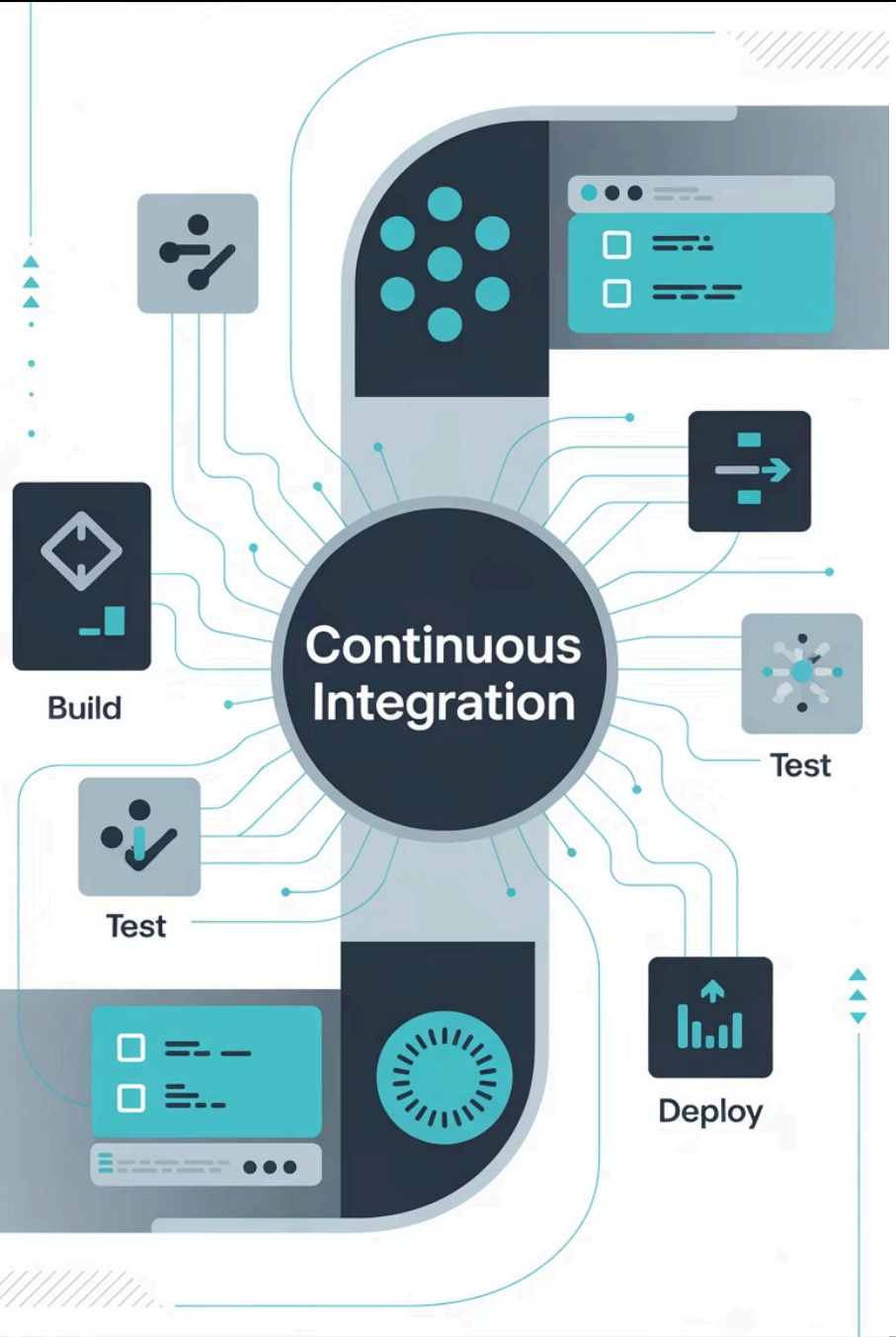
Monitors model accuracy across time horizons

Hyperparameter Search

Explores optimal configurations dynamically

Dynamic Timeout Redistribution

Allocates computational resources intelligently



Automated Production Pipeline

The entire workflow operates autonomously through `.github/workflows/main.yml`, embodying DevOps principles for research reproducibility.

1. **Data Fetching**
   Automated CVE retrieval
2. **Model Training**
   Ensemble optimization
3. **Forecasting**
   Prediction generation
4. **Deployment**
   Website updates

Overall CVE Forecasting Results

Our ensemble approach achieves remarkable accuracy using standard time-series metrics:

0.04%

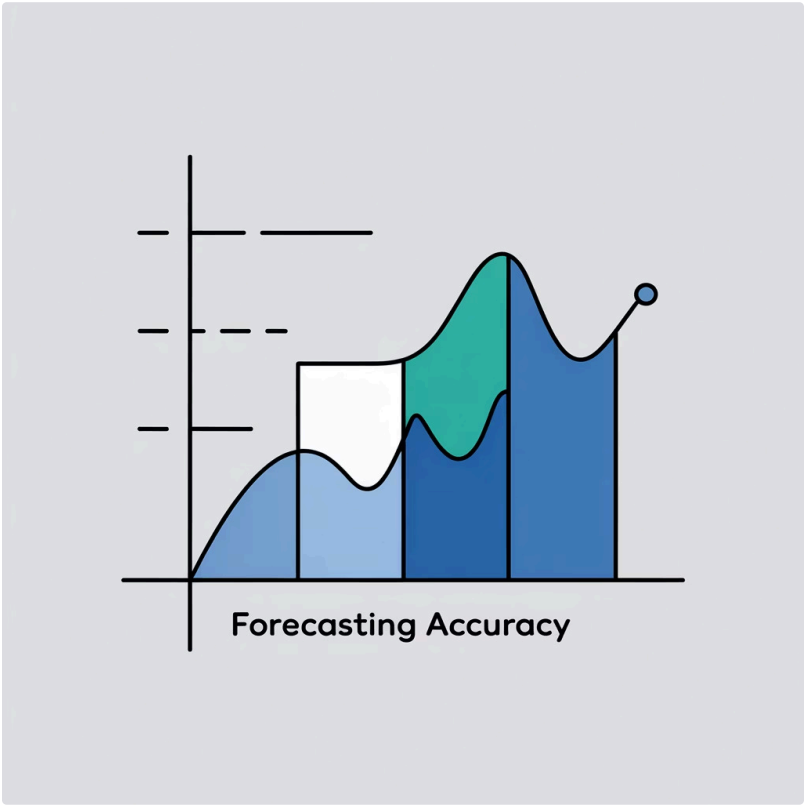
MAPE

Mean Absolute Percentage Error demonstrates strong predictive capability

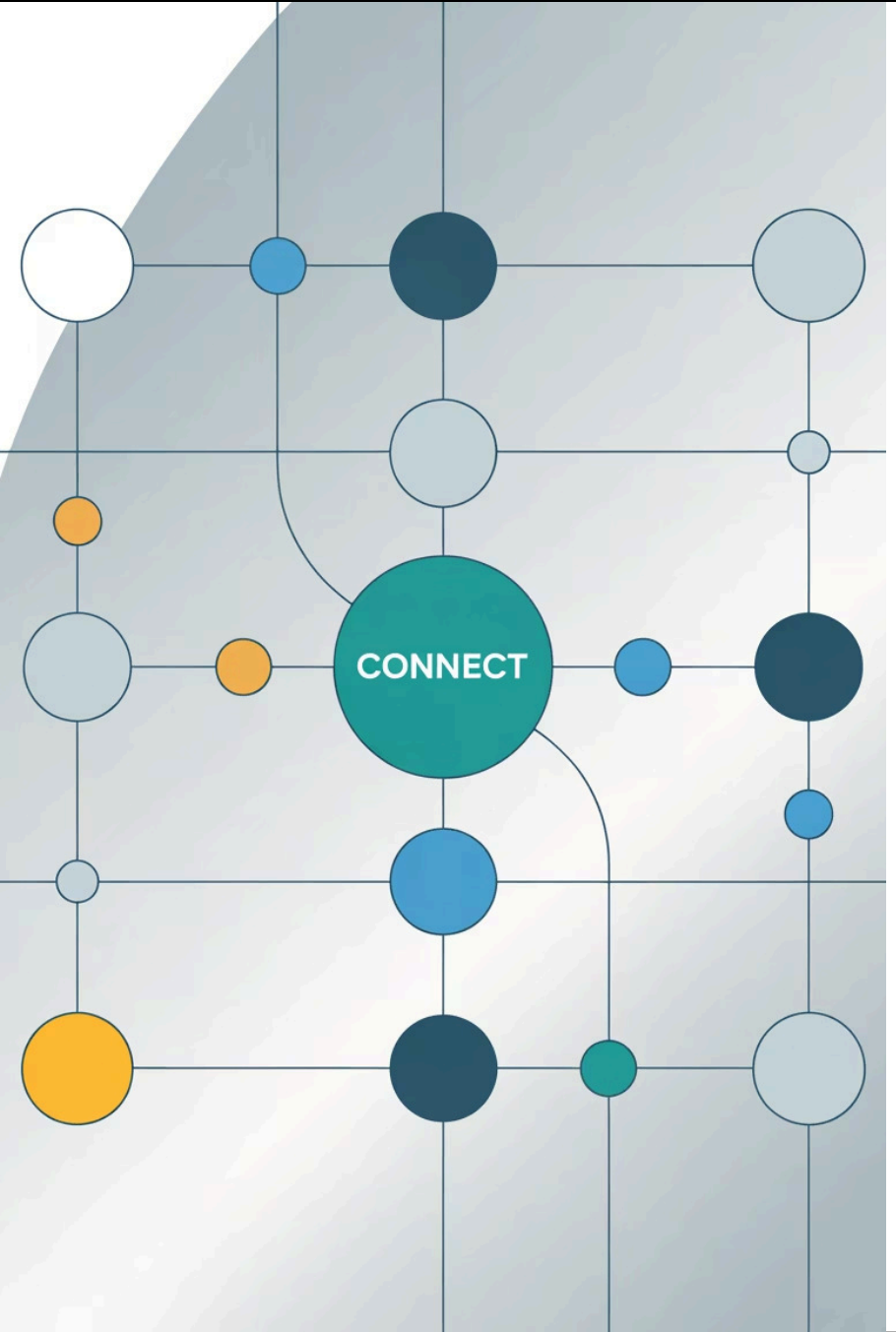
1.37

RAE

Relative Absolute Error shows consistent performance across time horizons



These results are based on data from the September 24th run.



CNA-Specific Forecasting Granularity

Individual Certificate Numbering Authority (CNA) forecasting represents a significant advancement in vulnerability intelligence granularity.

This capability enables organizations to anticipate disclosure patterns from specific vendors, transforming strategic security planning from reactive to predictive.

The `web/cna_forecast.html` interface provides actionable insights for targeted vulnerability management strategies.

Open Floor Discussion

Leveraging the collective expertise of vulnerability forecasting researchers to guide project evolution and community-driven development.

Future Research Directions

CVSS Score Forecasting

Should we predict average scores or full distribution patterns? Each approach offers distinct strategic advantages for risk assessment.

CVSS v4 Adoption

Forecasting adoption rates and impact on vulnerability streams could inform industry transition strategies.

CWE-Specific Forecasting

Predicting volumes for specific weaknesses like SQL injection or XSS enables targeted defense prioritization.

Broader Security Applications

Time-Series Opportunities

- Threat intelligence volume prediction
- Security incident forecasting
- Patch deployment timeline optimization
- Security tool effectiveness trends

The methodology's flexibility opens new research avenues across the security domain.

Collaborative Innovation

CVEforecast embodies open-source principles, inviting the vulnerability research community to contribute, validate, and extend our predictive capabilities.

Together, we can transform cybersecurity from reactive defense to proactive intelligence.