Reality checks for evaluation practices

in Time Series Forecasting

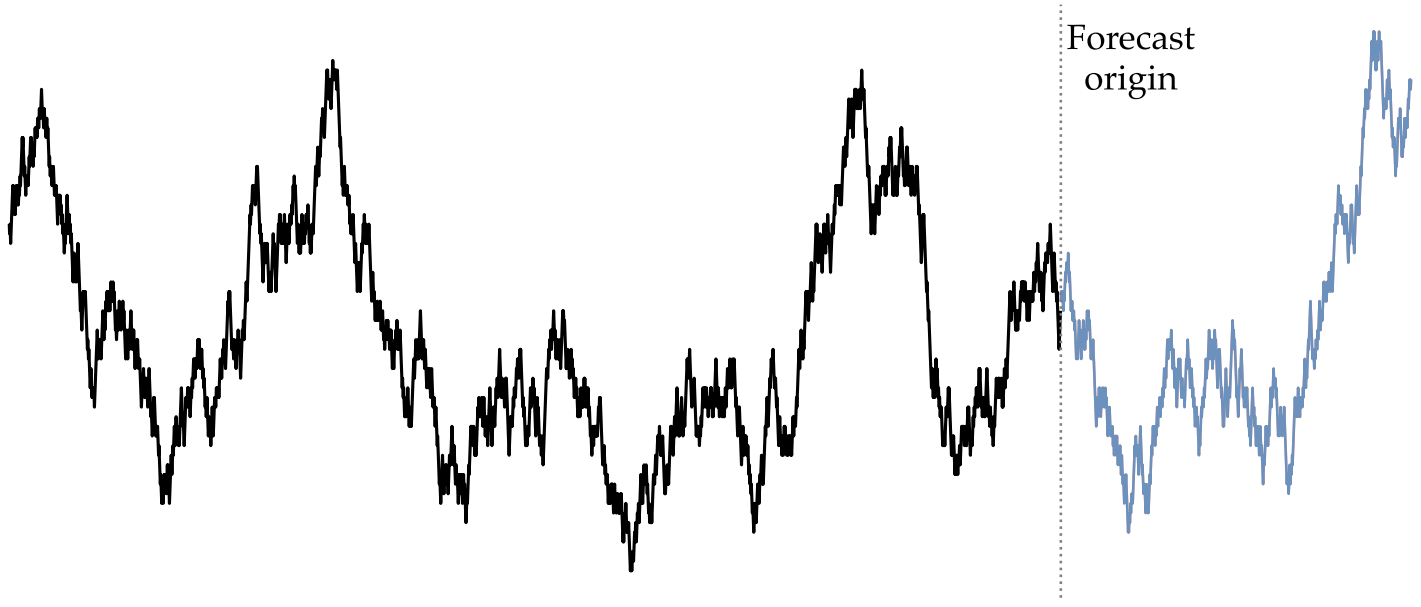
Vitor Cerqueira

vcerqueira@fe.up.pt

Faculdade de Engenharia, Universidade do Porto

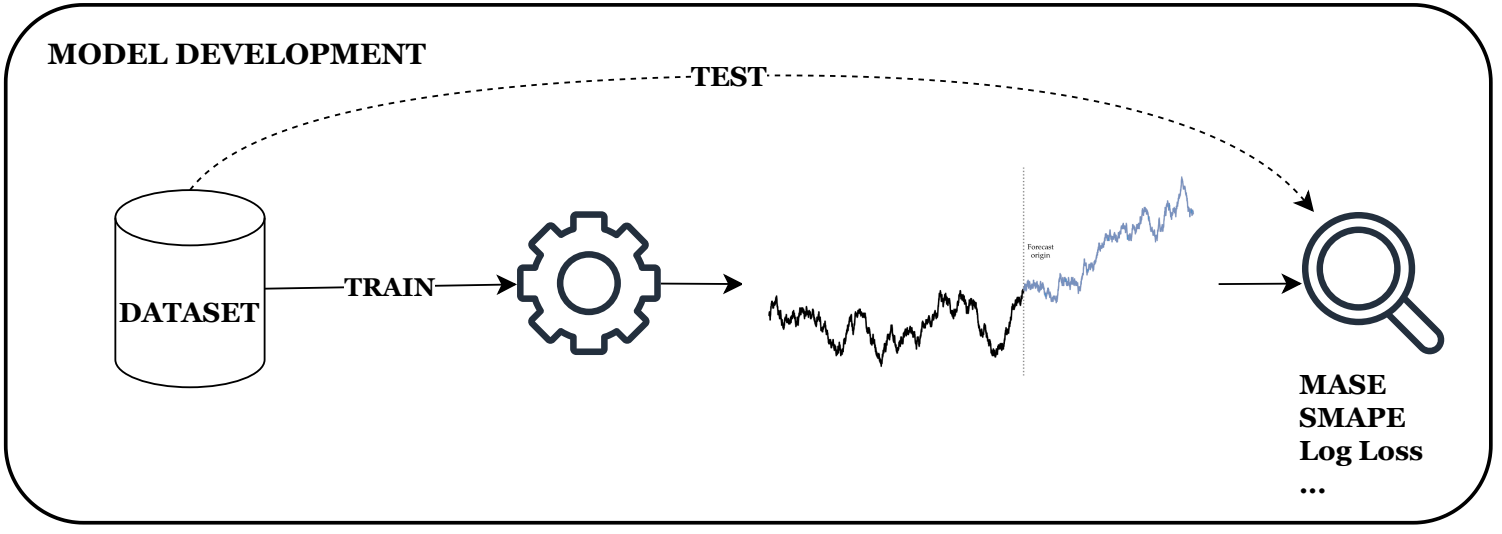
September 15, 2025

Time Series Forecasting



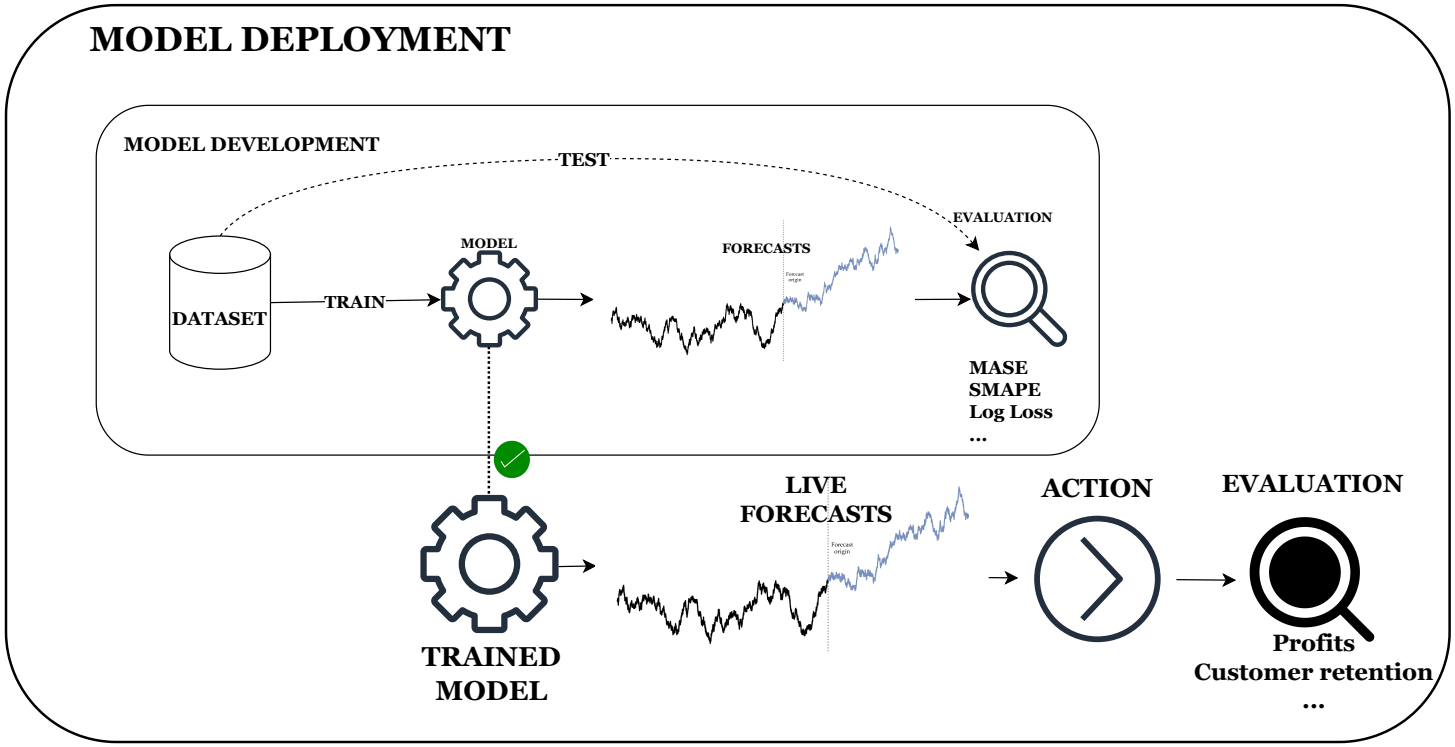
- Forecasting: Predicting future observations based on historical data
- ... How do we evaluate and compare different approaches?

Typical Model Development Workflow



- Model development
  - Focus on prediction accuracy metrics
  - ... and what happens after deployment?

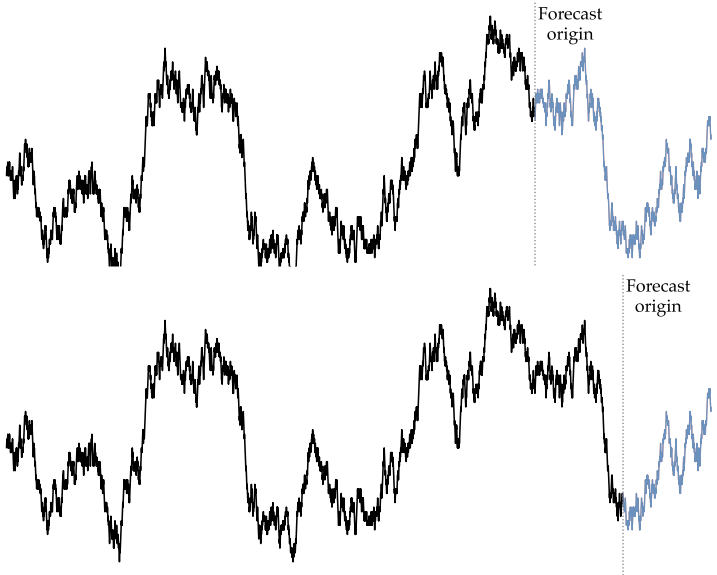
Typical Model Development Workflow

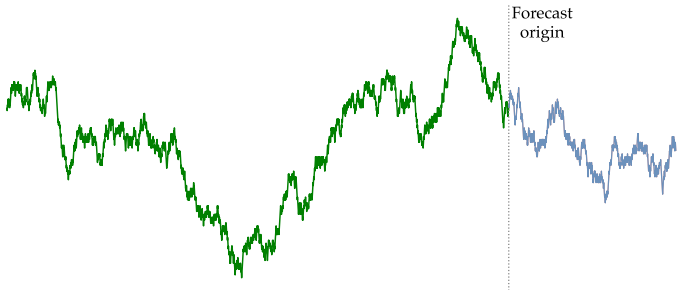


- Decision-making process is not considered during development

Outline

Myopic Evaluations





- Issue # 1: Myopic Evaluations. Metrics are typically averaged across all testing samples.
  - Multiple time steps
  - Multiple forecasting horizons
  - Across collections of time series
  - And other conditions (e.g. stationarity)
- Convenient, but myopic

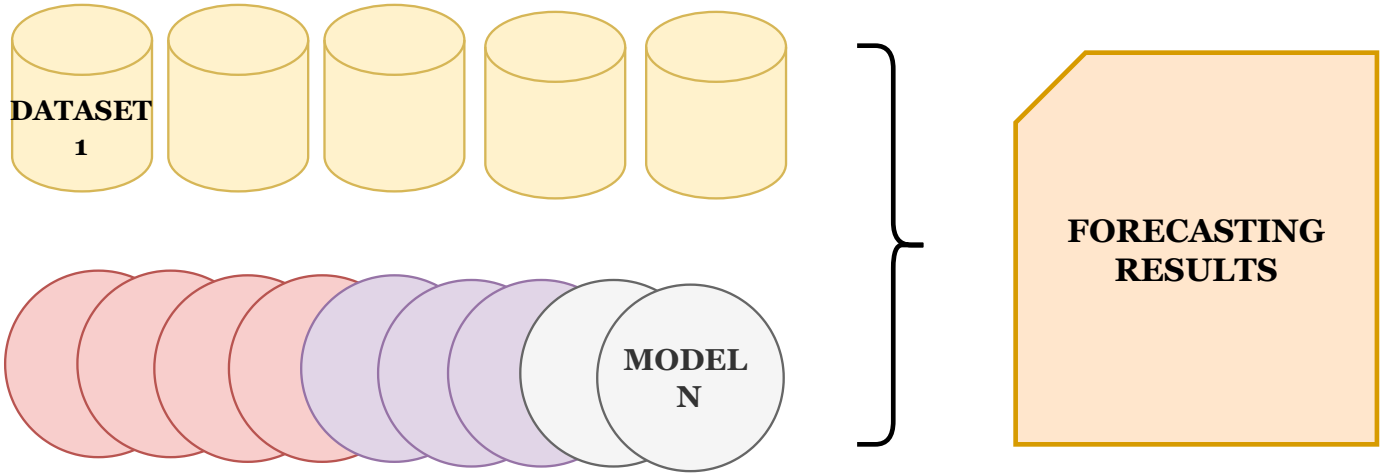
Averages of Averages of Averages...

Average performance is a convenient way to compare models and select the best performer

- ... But dilutes important bits of information.

- What are the strengths and limitations of a model?
- In which conditions does it perform poorly?
- Are performance differences significant or systematic?

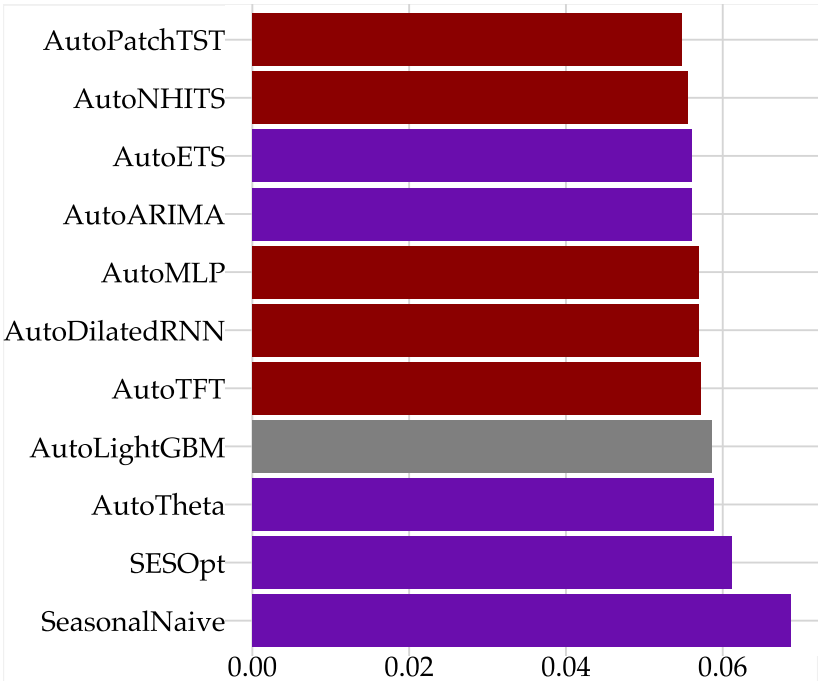
Experiments with Aspect-based Evaluation



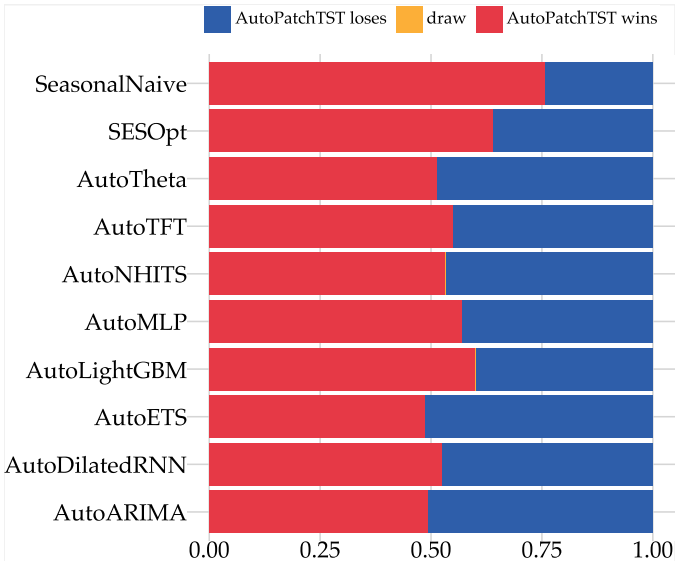
- +75k time series, + 14m data points
  - Monthly and quarterly frequencies
- Different methods, incl. neural nets, classical approaches
- Results controlled for several aspects

Cerqueira, V., Roque, L., & Soares, C. (2025). ModelRadar: Aspect-based Forecast Evaluation. Machine Learning (accepted).

Experiments with Aspect-based Evaluation



(a) SMAPE results



(b) Win/loss probability

- Figure: Average SMAPE (a) and (b) probability of PatchTST outperforming other approaches across all time series
- PatchTST performs overall best
  - ... but superiority depends on several conditions

Cerqueira, V., Roque, L., & Soares, C. (2025). ModelRadar: Aspect-based Forecast Evaluation. Machine Learning (accepted).

Controlling for Forecasting Horizon

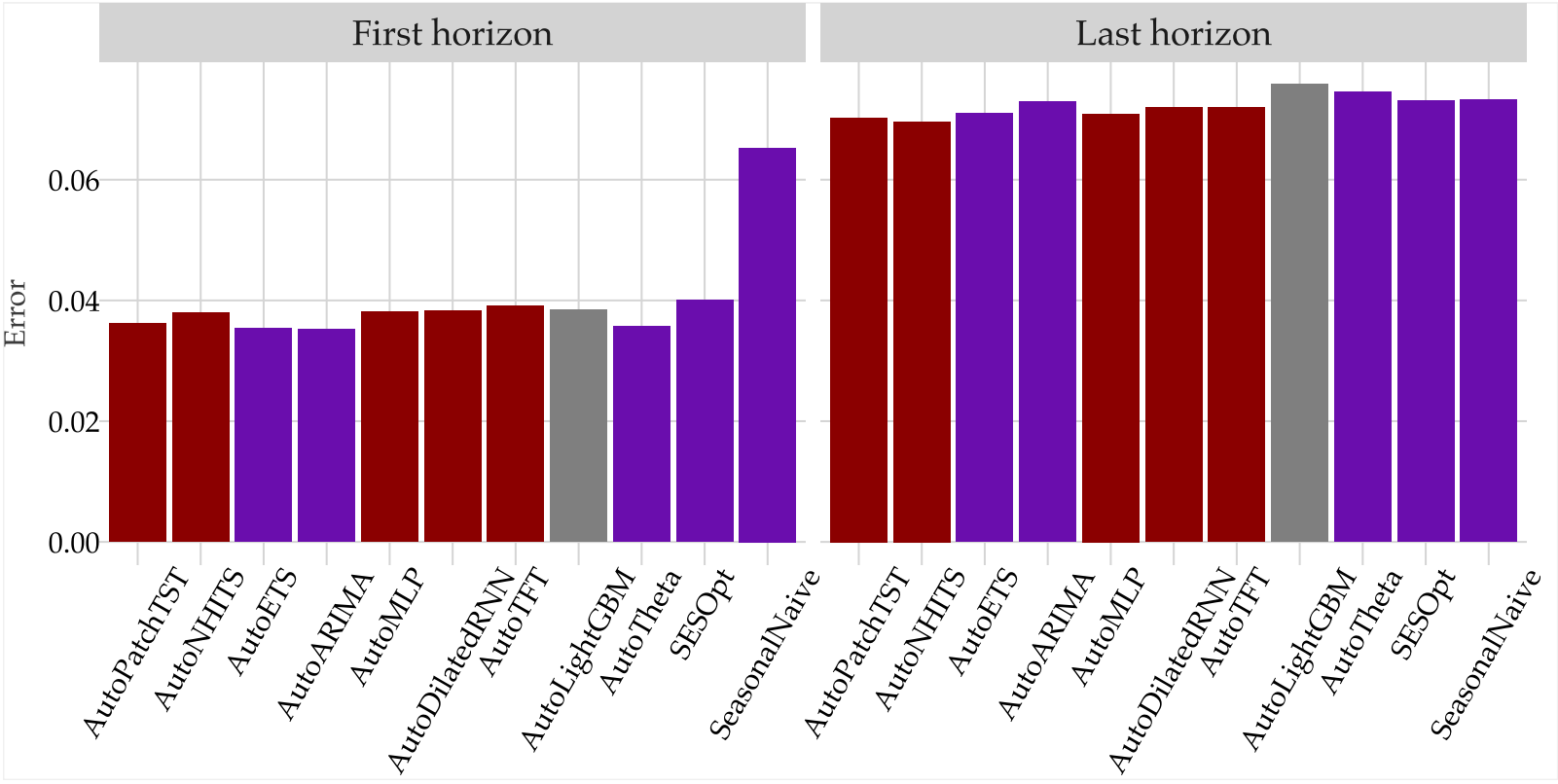


Figure: SMAPE scores of each model controlling for horizon condition.

- Neural networks only outperform classical approaches for multi-step ahead forecasting

Controlling for Stationarity

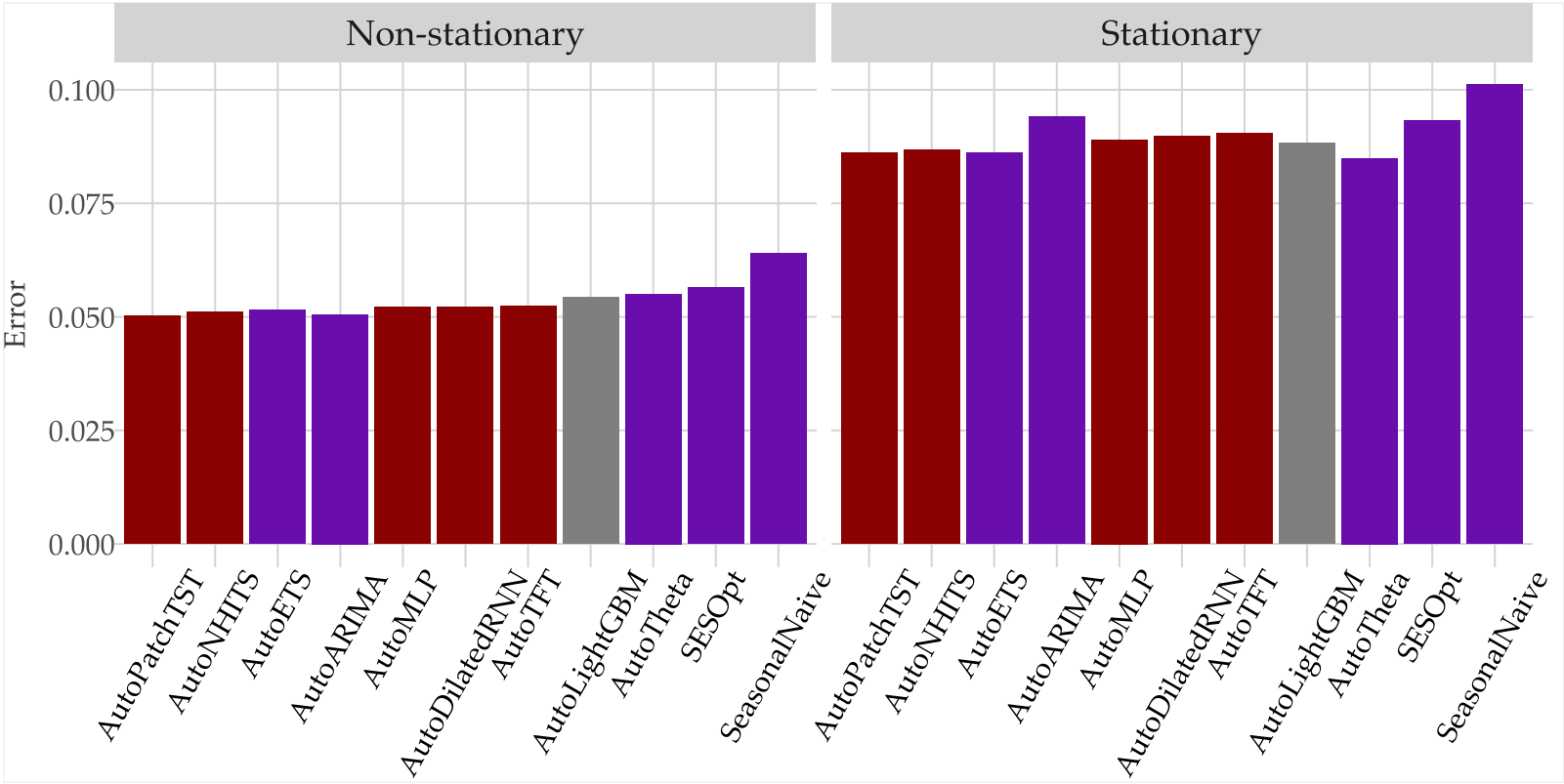
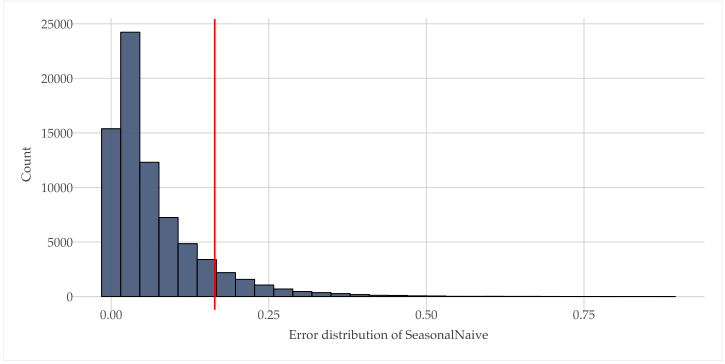


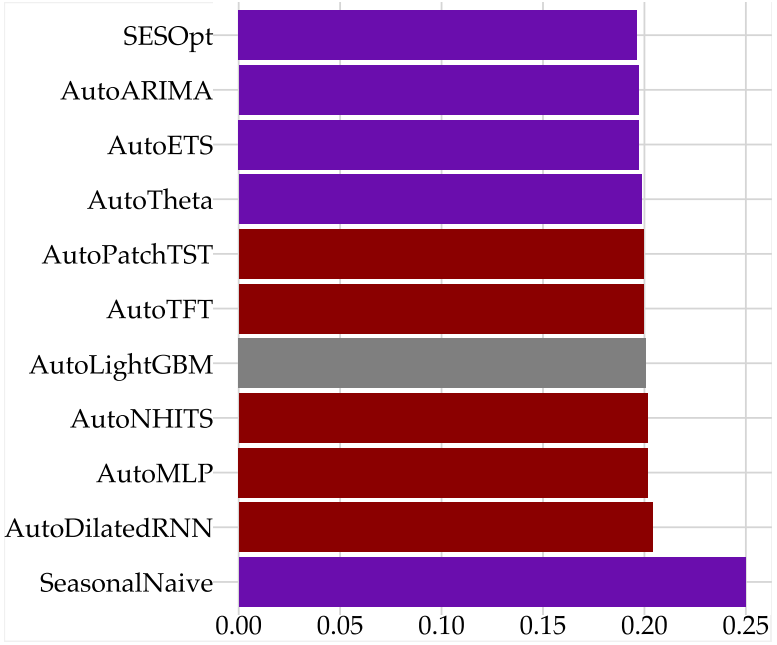
Figure: SMAPE scores of each model controlling for stationarity.

- In stationary time series, exponential smoothing methods are better than others

On *Hard* Problems



(a) SMAPE distribution



(b) Average SMAPE

- *Hard* problem: Worse 10% cases for the seasonal naive
- Classical methods outperform neural networks and gradient boosting

Model Radar

Figure: Rank of each model across different dimensions. Values far from the center represent better rank.

Model Radar

`github.com/vcerqueira/modelradar`

Cerqueira, V., Roque, L., & Soares, C. (2025). ModelRadar: Aspect-based Forecast Evaluation. Machine Learning (accepted).

Poor Benchmarking Practices

- Issue # 2 : Poor Benchmarking Practices

- Poorly selected baseline and reference methods for comparison, including unfair tuning efforts
- Limited number of datasets

Reference method selection

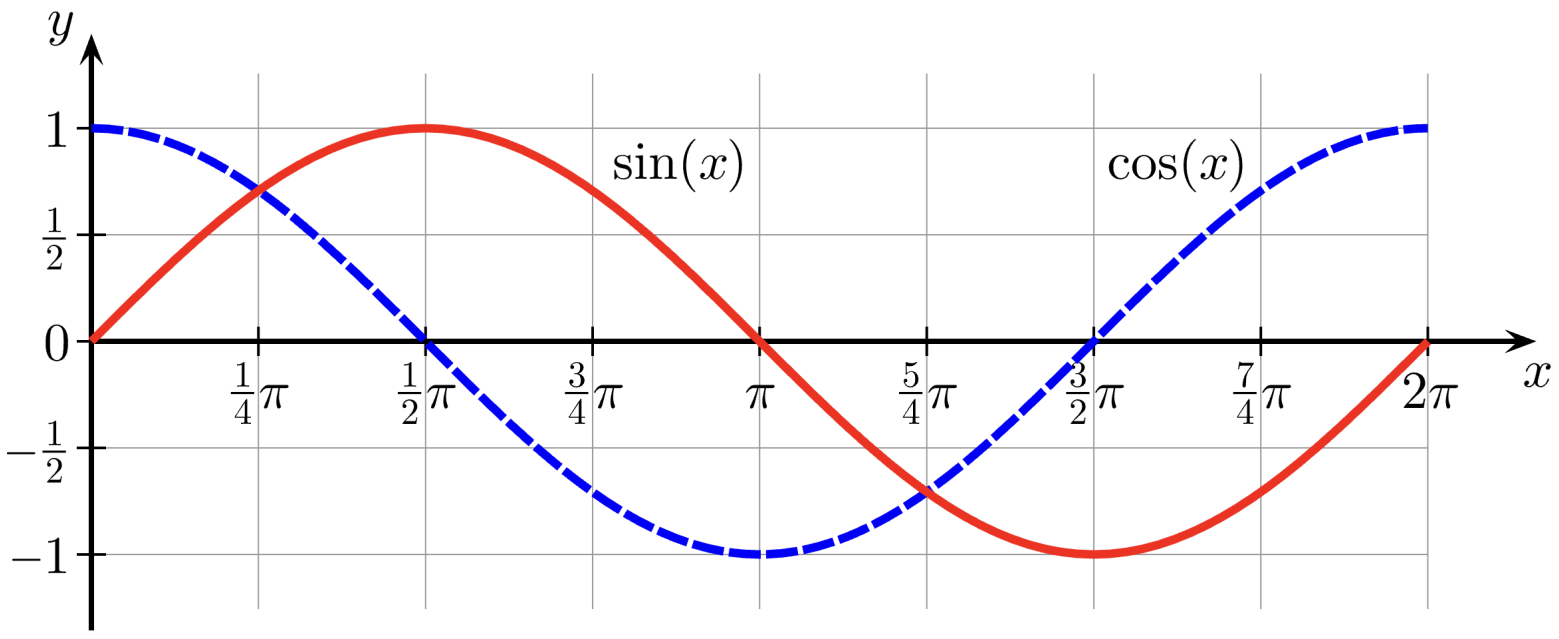
Table 4 Results of the DHR-ARIMA model along with Informer and the other benchmarks on the univariate forecasting task in the work by Zhou et al. (2021)

Model	$ETTh_1$ (720)		ECL (960)	
	MSE	MAE	MSE	MAE
Informer	0.269	0.435	0.582	0.608
Informer†	0.257	0.421	0.594	0.638
LongTrans	0.273	0.463	0.624	0.645
Reformer	2.112	1.436	7.019	5.105
LSTMa	0.683	0.768	1.545	1.006
DeepAR	0.658	0.707	0.657	0.683
ARIMA	0.659	0.766	1.370	0.982
Prophet	2.735	3.253	6.901	4.264

- Informer: Transformer-based neural net
  - almost 7k citations in $<$ 5 years
- Significantly outperforms ARIMA (among others)

Hewamalage, H., Ackermann, K., & Bergmeir, C. (2023). Forecast evaluation for data scientists: common pitfalls and best practices. Data Mining and Knowledge Discovery, 37(2), 788-832.

Reference method selection



https://en.wikipedia.org/wiki/File:Sine_cosine_one_period.svg

- DHR-ARIMA: Adding a few sine and cosine waves as covariates...

Reference method selection

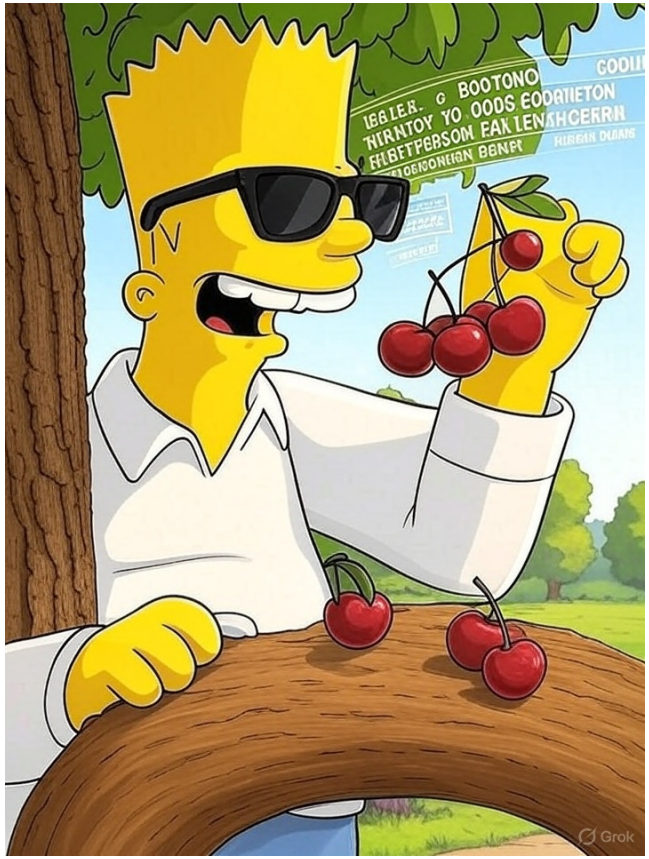
Table 4 Results of the DHR-ARIMA model along with Informer and the other benchmarks on the univariate forecasting task in the work by Zhou et al. (2021)

Model	$ETTh_1$ (720)		ECL (960)	
	MSE	MAE	MSE	MAE
Informer	0.269	0.435	0.582	0.608
Informer†	0.257	0.421	0.594	0.638
LongTrans	0.273	0.463	0.624	0.645
Reformer	2.112	1.436	7.019	5.105
LSTMa	0.683	0.768	1.545	1.006
DeepAR	0.658	0.707	0.657	0.683
ARIMA	0.659	0.766	1.370	0.982
Prophet	2.735	3.253	6.901	4.264
DHR-ARIMA	**0.140**	**0.297**	**0.433**	**0.499**

- Informer: Transformer-based neural net
  - almost 7k citations in $<$ 5 years
- DHR-ARIMA: ARIMA + Fourier features

Hewamalage, H., Ackermann, K., & Bergmeir, C. (2023). Forecast evaluation for data scientists: common pitfalls and best practices. Data Mining and Knowledge Discovery, 37(2), 788-832.

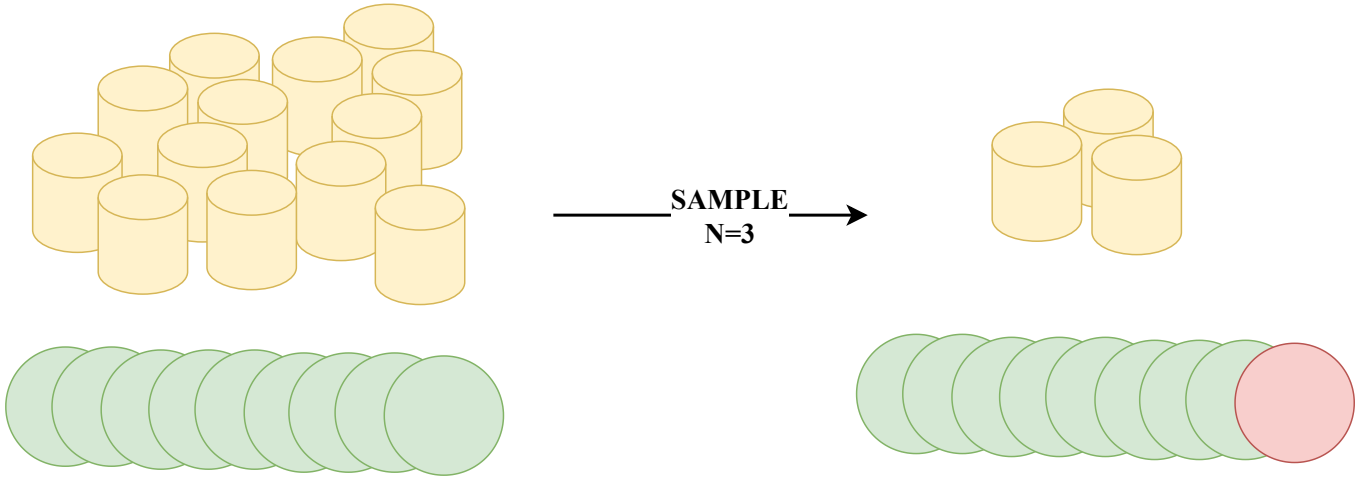
Dataset selection bias



Created by Grok

- Cherry-picking datasets to make your model shine

Cherry-picking Datasets



- Extensive experiments with several benchmark datasets and algorithms
- Systematically sample N datasets and analyse Top-*k* datasets for each algorithm

Roque, L., Cerqueira, V., Soares, C., & Torgo, L. (2025, April). Cherry-picking in time series forecasting: How to select datasets to make your model shine. In Proceedings of the AAAI Conference on Artificial Intelligence.

Cherry-picking Datasets

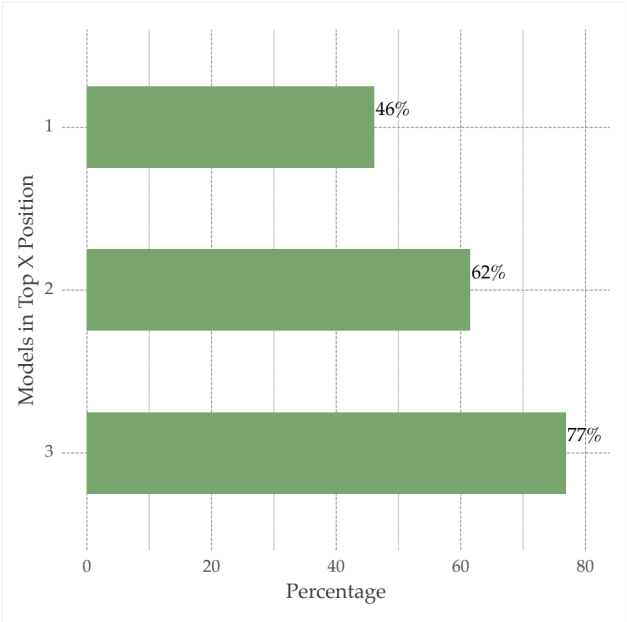


Figure 3: Percentage of models that could be reported as top 1, 2, and 3 performers based on an experimental setup of 4 datasets.

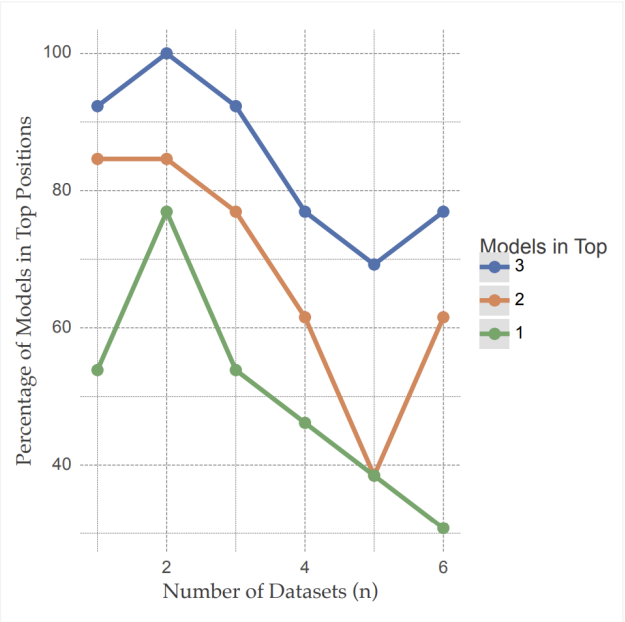
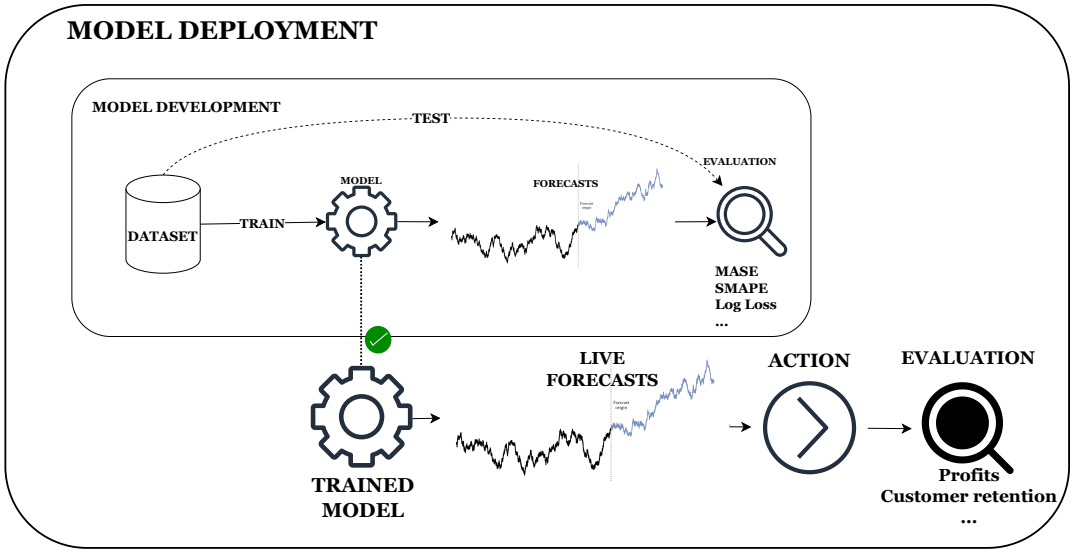


Figure 4: Breakdown of the percentages for top 1, 2, and 3 positions across different numbers of datasets.

- with 4 datasets (most common setting):
  - 46% of models could be the best (77% in top 3)
- with 6 datasets
  - significant reduction in the selection of wrong winner

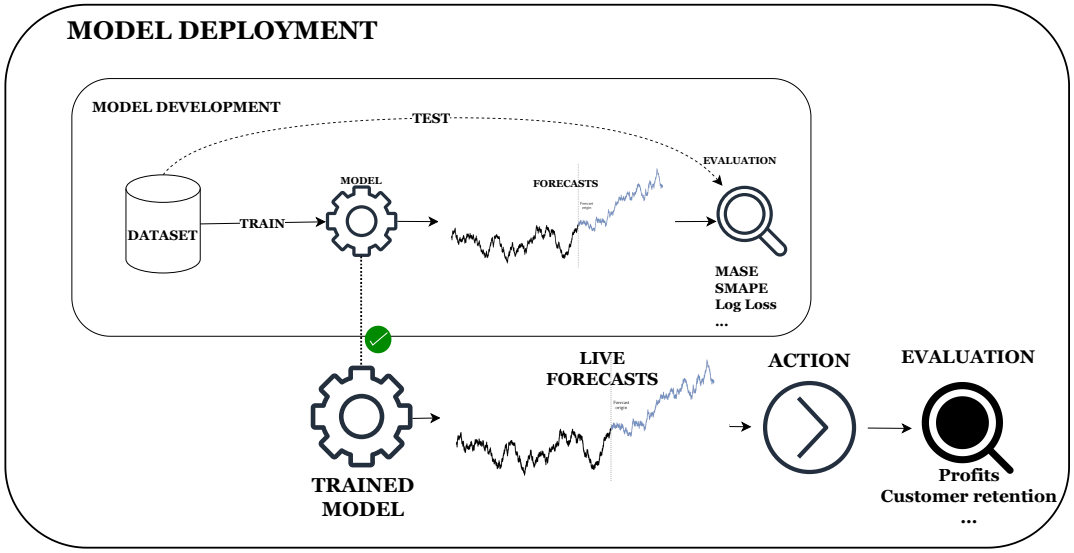
Roque, L., Cerqueira, V., Soares, C., & Torgo, L. (2025, April). Cherry-picking in time series forecasting: How to select datasets to make your model shine. In Proceedings of the AAAI Conference on Artificial Intelligence.

Ignoring Impact of Forecasts



- Issue # 3: Ignoring impact of forecasts on decision-making process.
- General assumption that accuracy is a good proxy for utility
  - Models are optimized for statistical accuracy of forecasts (beliefs about future events)
  - No consideration on the impact of forecasts on decision-making processes (payoff)

One Does Not Eat Forecasts



- Increasing accuracy does not necessarily lead to better payoffs
  - Gap between forecasts and decision-making
  - Non-uniform costs/benefits

Taleb, N.N.: On the statistical differences between binary forecasts and real-world payoffs. International Journal of Forecasting

Yardley, L., Petropoulos, F.: Beyond error measures to the utility and cost of the forecasts. Foresight: the International Journal of Applied Forecasting

Case study: Booking

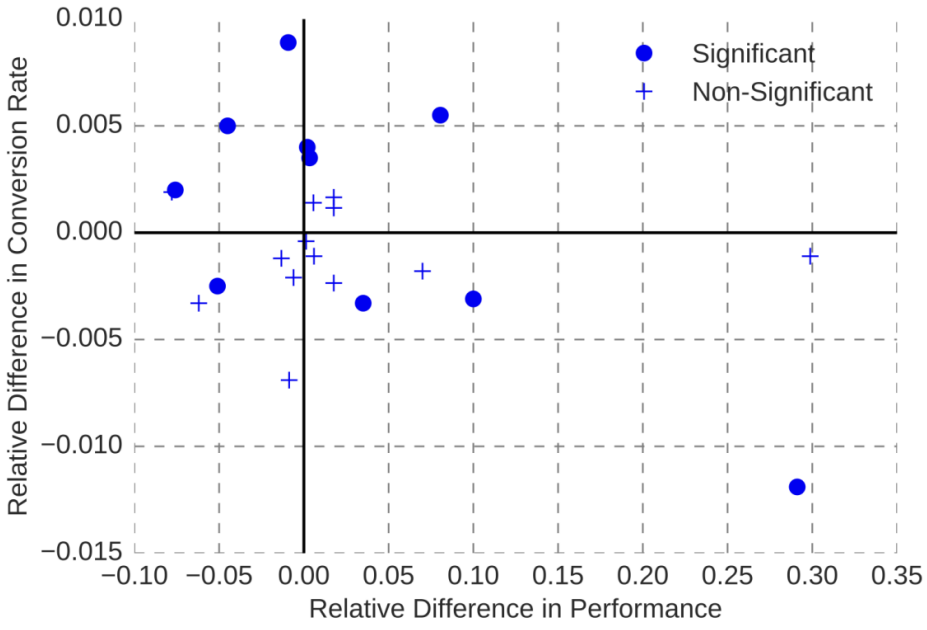


Figure 4: Relative difference in a business metric vs relative performance difference between a baseline model and a new one.

- No correlation between offline model performance (AUC) and utility (Conversion rate)
- "*Offline model performance metrics are only a health check, to make sure the algorithm does what we want to.*"

Bernardi, L. et al.. 150 successful machine learning models: 6 lessons learned at booking.com. KDD'19.

Other examples

- Finance: Stock-return prediction models
  - "...statistical performance is a very weak predictor of economic performance."

    Cenesizoglu, T. and Timmermann, A., 2012. Do return prediction models add economic value? Journal of Banking and Finance

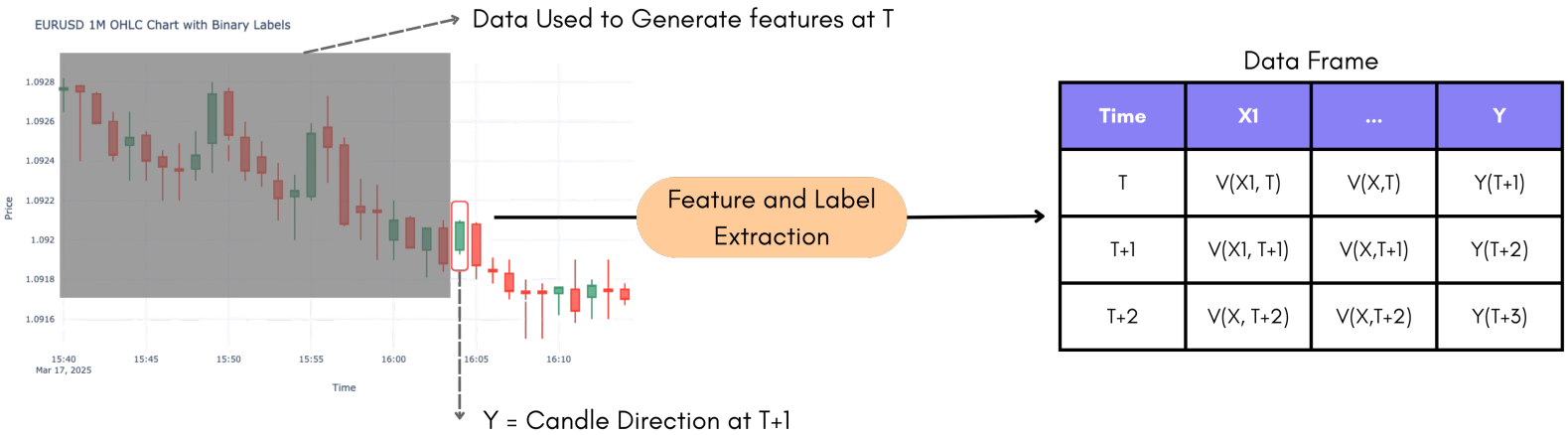
- Betting
  - "...forecasting accurately and forecasting profitably are different tasks and should be treated as such."

    Wunderlich, F. and Memmert, D., 2020. Are betting returns a useful measure of accuracy in (sports) forecasting? International Journal of Forecasting

- Meteorology, Supply Chain, etc.

Yardley, L., Petropoulos, F.: Beyond error measures to the utility and cost of the forecasts. Foresight: the International Journal of Applied Forecasting

Our Case Study: Binary Options Trading



Time	X1	...	Y
T	V(X1, T)	V(X,T)	Y(T+1)
T+1	V(X1, T+1)	V(X,T+1)	Y(T+2)
T+2	V(X, T+2)	V(X,T+2)	Y(T+3)

- High-frequency trading simulation
  - Binary probabilistic forecasting
  - Risk-Adjusted Position Sizing
  - etc..
- Several classifiers evaluated with different metrics

Santos, J., Soares, C., Branco, P. & Cerqueira, V. (2025). Does Improving Forecasting Accuracy Also Improves Financial Utility?

Our Case Study: Binary Options

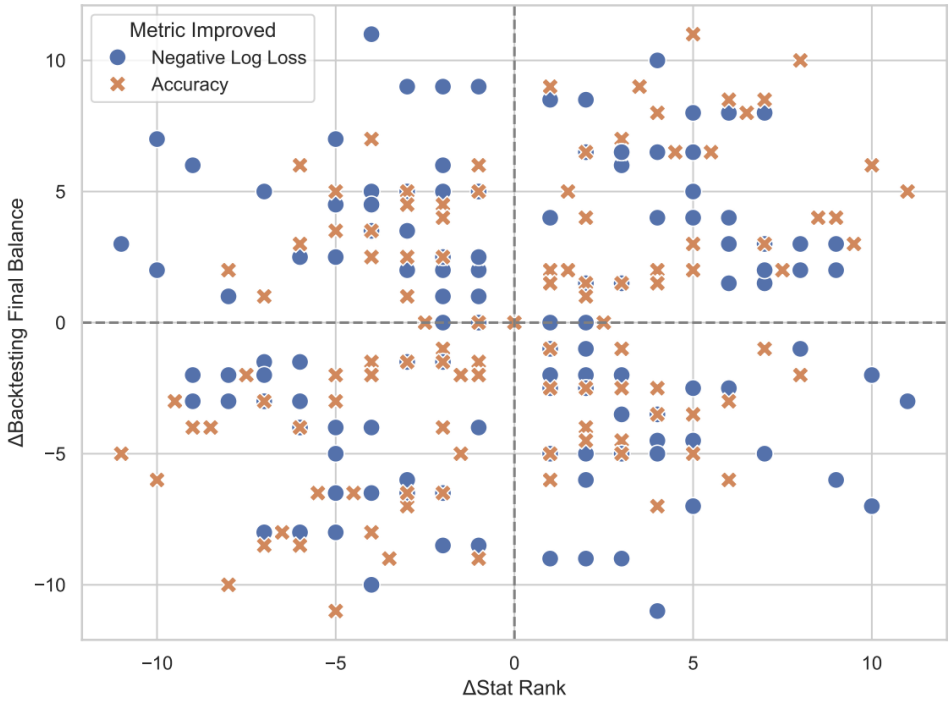


Fig. 2: Comparing rank changes in classification accuracy vs change in account final balance. Each point represents a model pair comparison.

- No relationship between accuracy/log loss and profit

Our Case Study: Binary Options

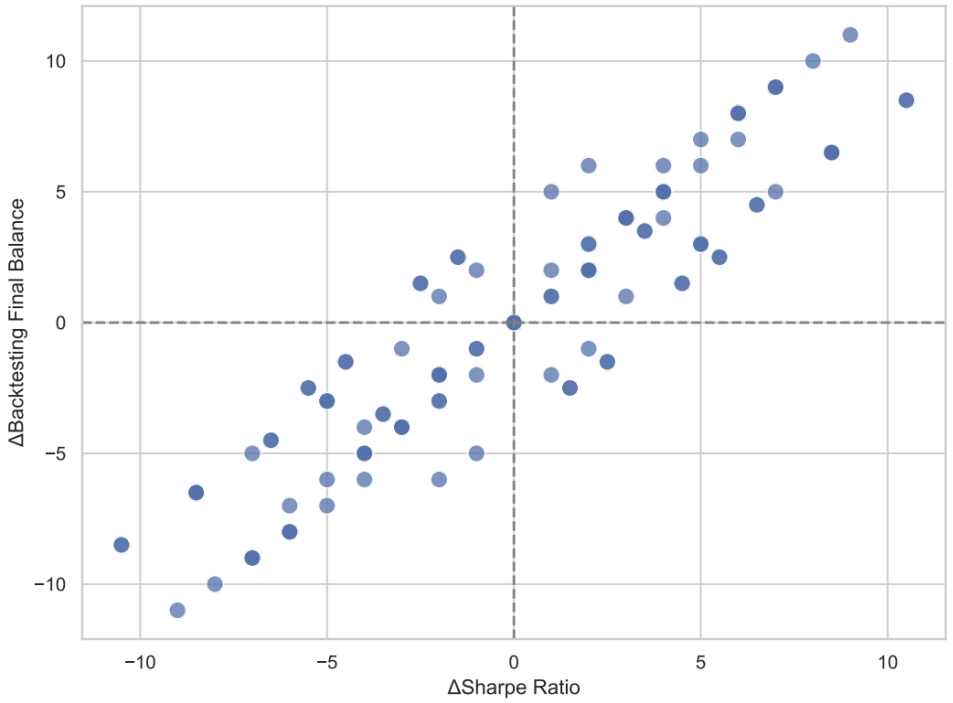
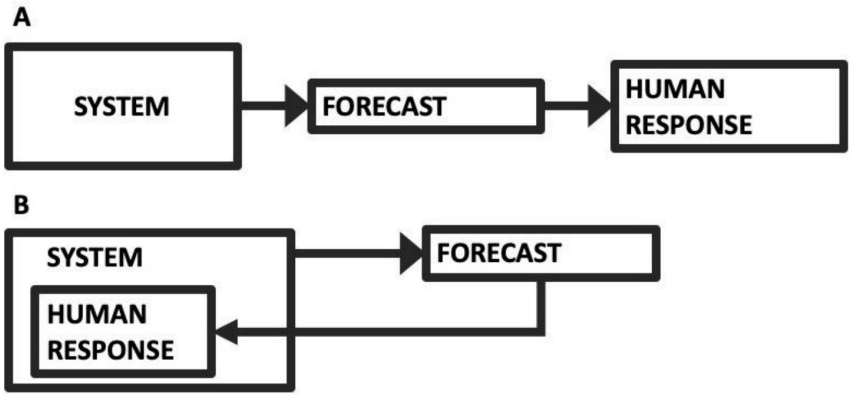


Fig. 3: Comparing rank changes in Sharpe ratio vs change in account final balance. Each point represents a model pair comparison.

- Sharpe ratio has clear relation with profit
  - Based on the trading decisions results from forecasts

Reflexivity



- Generic assumption that a forecast bears no impact on the underlying system
- Not true in many domains, e.g. social sciences
  - Actions based on forecasts can make them less or more accurate
  - Self-fulfilling or self-defeating prophecies

Record, N.R.; Pershing, A.J. Facing the Forecaster's Dilemma: Reflexivity in Ocean System Forecasting. Oceans 2021, 2, 738-751

What can we do?

- Incorporating forecast added value evaluation
  - How will the forecasts be used?
  - How will forecasts impact the decision at hand?
  - What are the expenses of forecast creation?
  - What are the costs associated with forecast error?

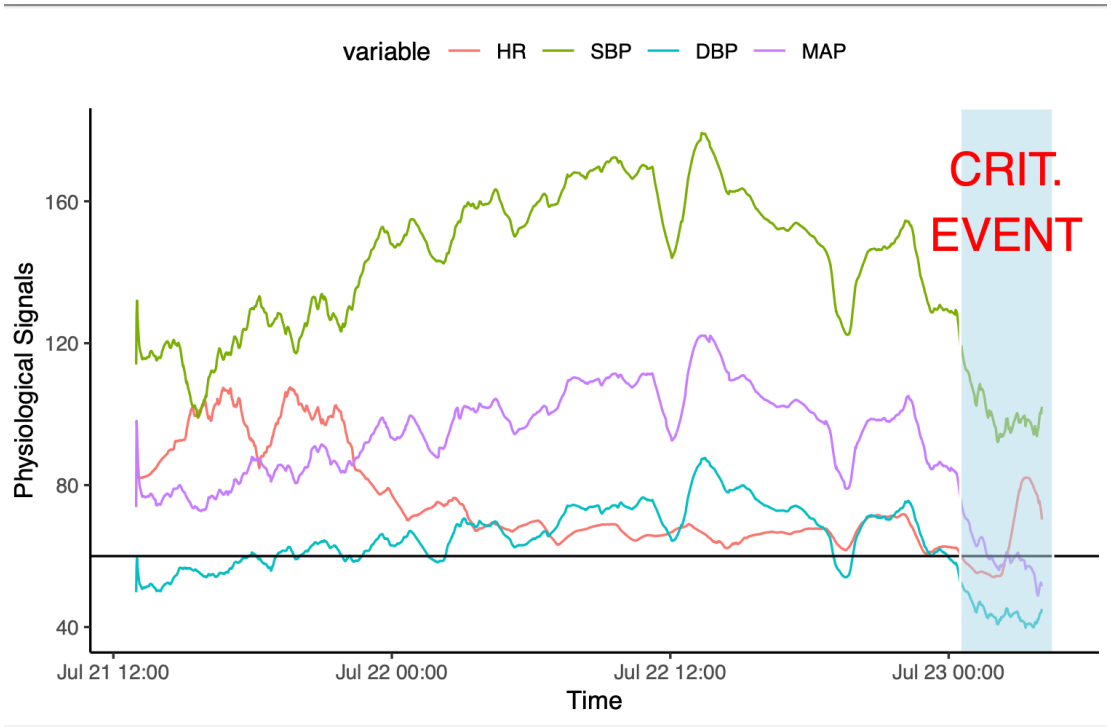
Yardley, L., Petropoulos, F.: Beyond error measures to the utility and cost of the forecasts. Foresight: the International Journal of Applied Forecasting

- Embed decision-making utility in model development

Donti, P., Amos, B., & Kolter, J. Z. (2017). Task-based end-to-end model learning in stochastic optimization. Advances in neural information processing systems, 30.

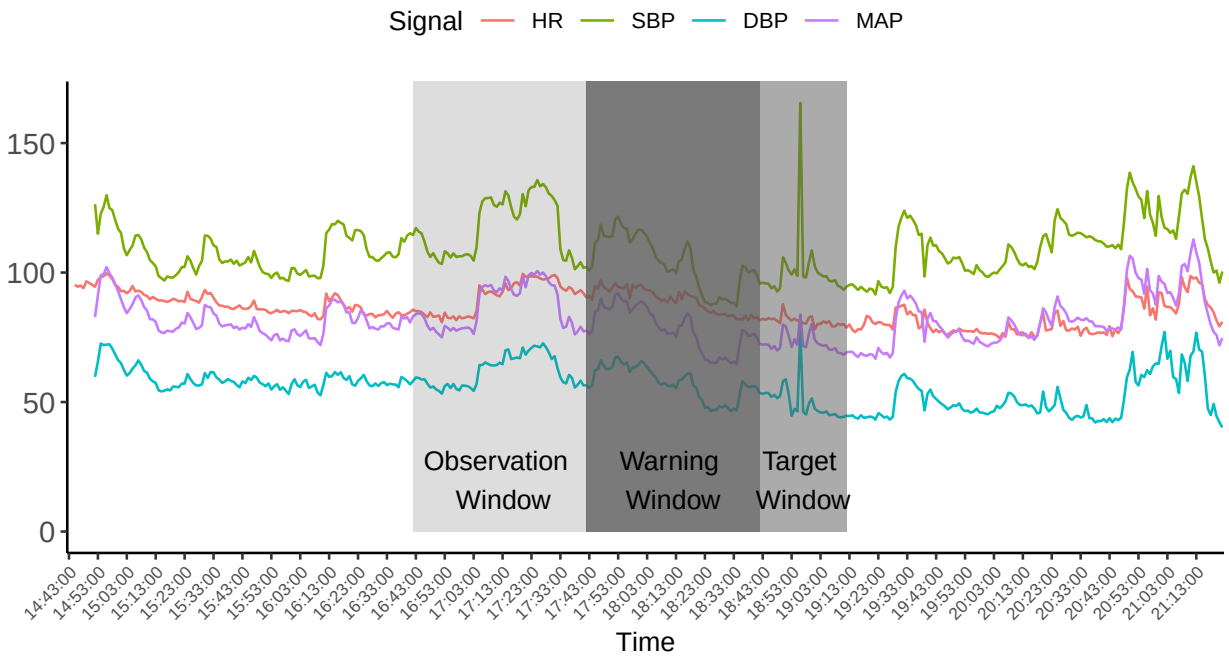
The Value of Time in Forecasting

- Issue # 4 : Ignoring the value of time when forecasting
- Particular example: Real-time anomaly detection



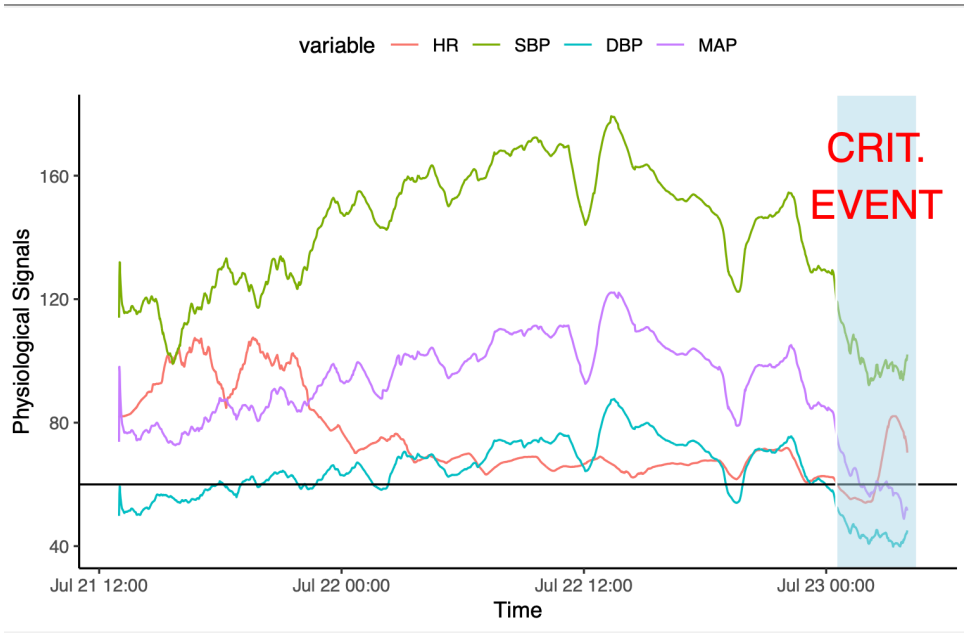
Cerqueira, V., Torgo, L., & Soares, C. (2023). Early anomaly detection in time series: a hierarchical approach for predicting critical health episodes. Machine Learning, 112(11), 4409-4430.

Typical Problem Formulation



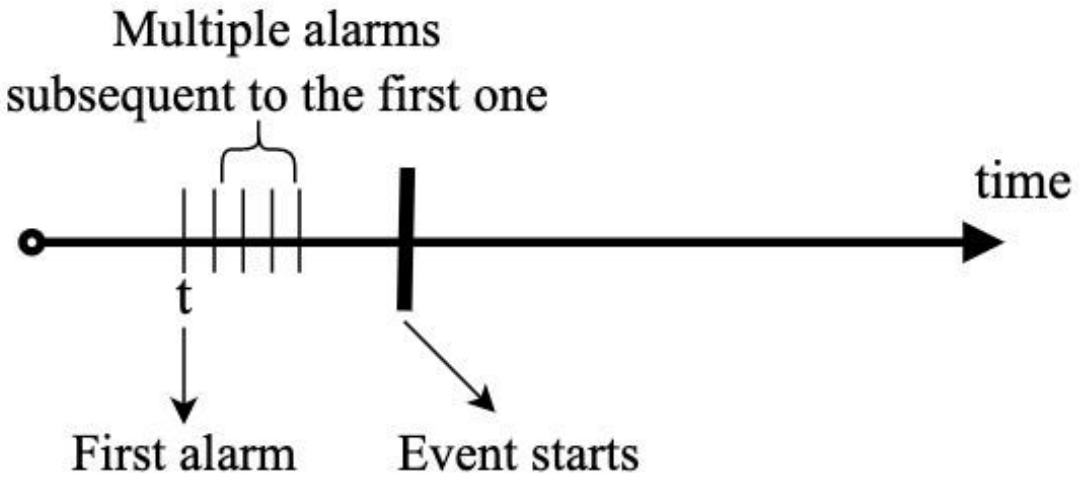
- Models typically evaluated **instance by instance**
  - AUC, log loss, F1, …
- But this approach ignores the value of timely predictions

The Value of Time in Forecasts



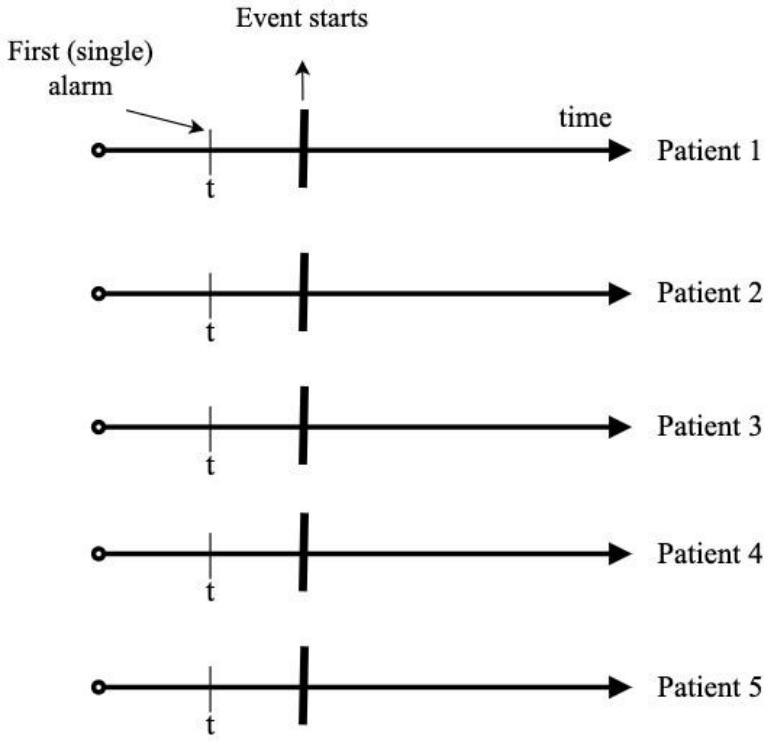
- Models are built to to classify each instance as positive or negative
- But the actual goal is to detect, in a timely manner, anomalous events

Value of consecutive alarms



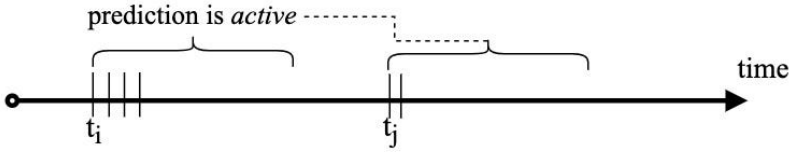
- Figure: Five consecutive alarms that correctly predict an event.
  - First alarm is useful because it elicits some action.
  - Subsequent ones add no information.

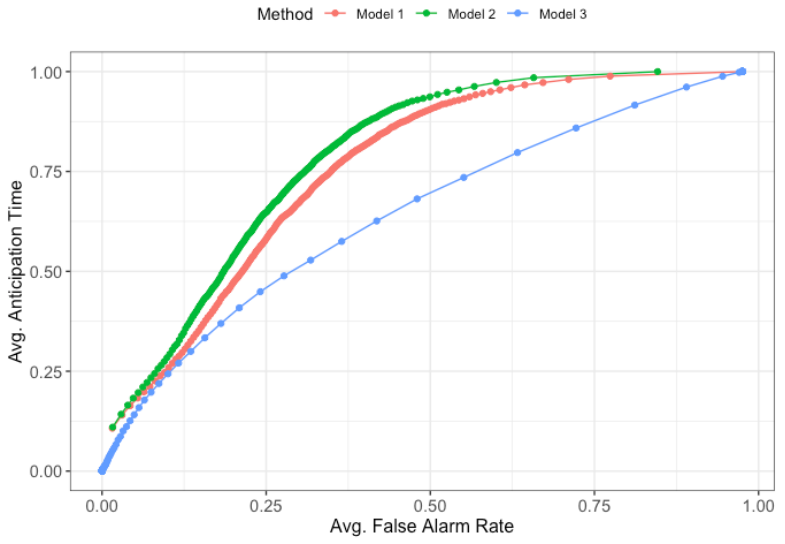
Value of consecutive alarms



- Figure: another case with 5 TP's, but in 5 different events
  - Standard metrics are blind to these nuances

Possible remedies…





- Metrics at the event-level
  - e.g. Event recall, reduced precision
- ROC analysis variations, e.g. AMOC (Activity Monitoring OC)
  - False alarm rate per unit of time vs event recall

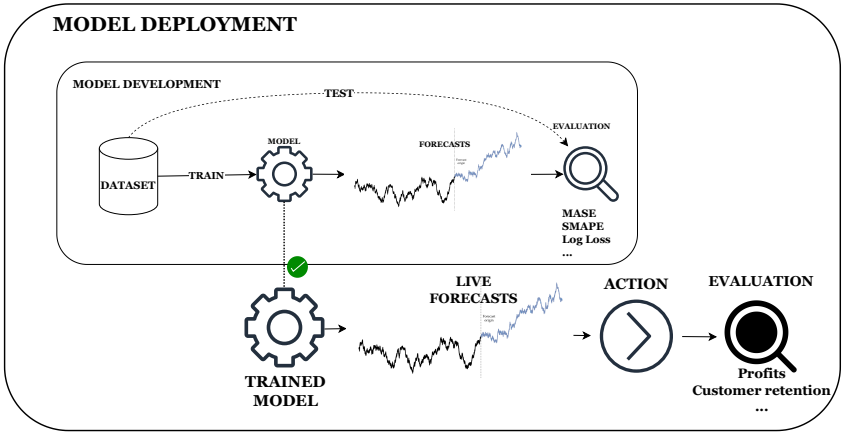
Fawcett, Tom, and Foster Provost. "Activity monitoring: Noticing interesting changes in behavior." KDD'99.
Weiss, Gary M., and Haym Hirsh. "Learning to Predict Rare Events in Event Sequences." KDD'98.

Verification delay

Time is also important for **verification** (getting labels)

- Forecasting: Clean historical data vs reality
  - Models are built on *clean* historical datasets
  - ... But inference is sometimes done using provisional measurements
- Data stream mining
  - Often assumes immediate feedback (no verification delay)
  - ... Labels can take a long time to arrive
    - Impact on change detection and adaptation

Wrapping up



1. Less myopic evaluations
2. More thorough benchmarks and comparisons
3. Understanding how forecasts are used in practice
4. Consider temporal constraints

Our Team @ Center for Responsible AI

Responsible AI Research | Privacy, Auditing, and Compliance Tools

Goals:

- Reliable empirical science of machine learning
- Responsible development of machine learning solutions

Q&A

Thank you!

`https://github.com/vcerqueira/`

PRR Plano de Recuperação e Resiliência
REPÚBLICA PORTUGUESA
Financiado pela União Europeia NextGenerationEU