

MANAGEMENT REVIEW

Management without theatre.

AI

DEMAND FORECASTING: WHAT MACHINE LEARNING CHANGED

A forecasting competition on Walmart's sales, most teams behind the simple benchmarks, what won twice, where the gains end, how a forecast and a hand adjustment are measured, and a review card.



22.4%

more accurate than the best statistical benchmark: the winner of the M5 competition on Walmart's sales, 2020.

MAKRIDAKIS ET AL., 2022

INSIDE

COVER STORY

Ten stores, 42,840 series

PAGE 03

WHAT THE RESEARCH SAYS

Where the gains end

PAGE 06

TOOL OF THE ISSUE

The forecast review card

PAGE 08

NO. 77 · AI

In this issue

A demand forecast decides what is ordered, how many people are on the roster and how much stock waits on the shelf. Machine learning promises better forecasts. This issue asks what the largest open forecasting competitions found it changed, and what it did not.

In the M5 competition of 2020, the winner forecast Walmart's sales 22.4% more accurately than exponential smoothing, yet only 7.5% of the teams beat that benchmark. In M4 and M5 alike, combinations won; machine learning won when it learned across many related series. The gains were large for total sales and small for one product in one store. Simple benchmarks tell whether a method helped, and four companies show when a planner's hand adjustment helped.

Stiven Janaqi EDITOR · STIVEN CATALYST

03 COVER STORY
Ten stores, 42,840 series

04 THE NUMBERS
Most did not beat the simple ones

05 THE MODEL
What won, twice

06 WHAT THE RESEARCH SAYS
Where the gains end

07 HOW IT IS MEASURED
Measure against the naive

08 TOOL OF THE ISSUE
The forecast review card

09 SOURCES
Sources and method

HOW TO READ THIS ISSUE

Figure

Every figure has its source and year at the foot of its page.

Our reading

Where the editors interpret rather than the research, it says so.

Practice

The steps and the card are proposals to try, not research results.

COVER STORY

Ten stores, 42,840 series

In 2020 the M5 competition gave forecasters more than five years of daily unit sales from Walmart: 3,049 products in ten stores in California, Texas and Wisconsin. The task was to forecast the next 28 days.

42,840

series in 12 levels: each product in each store, summed by department, category, store and state

30,490

forecasts asked for at the lowest level, one product in one store; the rest summed up from them

The winning team was 22.4% more accurate than the best statistical benchmark, exponential smoothing. Only the top five passed 20%; all of the top 50 passed 14%. The winner averaged, with equal weights, many LightGBM models, a method built from decision trees, each trained across all the series of a store, a store category or a store department.

OUR READING

Part of the gain came from modelling seasonality and using outside information such as prices; the rest, the organisers think, from learning across thousands of related series at once.

Sources: Spyros Makridakis et al., International Journal of Forecasting 38(4), The M5 competition: Background, organization, and implementation, 2022 (via Abstract; the intermittency shares from a later paper); Spyros Makridakis et al., International Journal of Forecasting 38(4), M5 accuracy competition: Results, findings, and conclusions, 2022 (via The authors' preprint, 2020)

Accuracy as WRMSSE: each series' error scaled by that of a naive forecast and weighted by its dollar sales. The results paper was read in the authors' 2020 preprint; the background paper only in its abstract.

THE NUMBERS

Most did not beat the simple ones

Next to the entries, the organisers ran benchmarks, from the naive forecast, which repeats the last value, to exponential smoothing. Of the teams in the final ranking, the share that beat each one:

TEAMS THAT BEAT EACH BENCHMARK, M5 ACCURACY, 2020

Naive forecast



Seasonal naive forecast



Exponential smoothing, the best benchmark



Had each team submitted its best attempt, 63.7%, 48.8% and 12.2% would have beaten them. The organisers suspect misleading validation scores: many teams did not pick the best method they had built.

OUR READING

A complex model is easy to build and hard to check. Without a simple benchmark next to it, nobody knows whether it helped.

Source: Spyros Makridakis et al., International Journal of Forecasting 38(4), 2022 (via The authors' preprint, 2020)

Final submissions. The organisers add that many teams behind exponential smoothing may have lost interest and stopped, so the 7.5% may be somewhat biased.

THE MODEL

What won, twice

Two years before M5, the M4 competition asked for forecasts of 100,000 time series. Its lessons and those of M5 point the same way, with one change.

M4, 2018

12 of the 17 most accurate methods were combinations

the winner, a hybrid of statistics and a neural network, about 10% better than a combination benchmark

six pure machine learning methods: none beat the combination, one beat the naïve

M5, 2020

LightGBM in most of the leading entries

combinations of many models

one model learning across many series

prices, promotions and events as inputs

OUR READING

Machine learning won where it had many related series and information beyond past sales, as in M5. Pure machine learning in M4 had not.

Sources: Spyros Makridakis et al., International Journal of Forecasting 34(4), 2018 (via Abstract in library records); Spyros Makridakis et al., International Journal of Forecasting 38(4), 2022 (via The authors' preprint, 2020)

The 10% in M4 is by sMAPE, a percentage error. In M5, exponential smoothing with promotions and special days was 6% more accurate than without; ARIMA with such inputs, 13%.

WHAT THE RESEARCH SAYS

Where the gains end

The 42,840 series sit on 12 levels, from all of Walmart's sales in the data down to one product in one store. The organisers compared the top methods with the best benchmark at each level.

AVERAGE GAIN OF THE TOP METHODS OVER THE BEST BENCHMARK, BY LEVEL (2020)



For one product in one store, 73% of the 30,490 series were intermittent: days with no sale, then a few. There, the organisers note, the standard methods for intermittent demand still held some value.

OUR READING

The forecast that fills the shelf is still the uncertain one. Plan stock, people and capacity for the error, not only for the number.

Sources: Spyros Makridakis et al., International Journal of Forecasting 38(4), M5 accuracy competition: Results, findings, and conclusions, 2022 (via The authors' preprint, 2020); Spyros Makridakis et al., International Journal of Forecasting 38(4), The M5 competition: Background, organization, and implementation, 2022 (via Abstract; the intermittency shares from a later paper)

Rounded averages, read in the authors' preprint of the results paper. At the three lowest levels the overall winner ranked 11th to 13th of the top 50.

[More in the essay](#)

High-volume days: the standard under pressure

HOW IT IS MEASURED

Measure against the naive

Rob Hyndman and George Athanasopoulos use four simple methods as benchmarks throughout their textbook; the simplest, the naive forecast, repeats the last value. Other methods are compared with them.

THREE MEASURES OF ERROR

01	MAE, mean absolute error	In units. Minimising it leads to the median; minimising RMSE, to the mean.
02	MAPE, mean absolute percentage error	Infinite or undefined on a day with zero sales, extreme near zero.
03	MASE, mean absolute scaled error	Below 1 when the forecast beats the naive forecast on past data.

In four supply-chain companies, over 60,000 forecasts, planners adjusted on average 75% of the system's forecasts. Large adjustments helped more; small ones often did harm. Upward ones helped far less often and more often went the wrong way.

OUR READING

Measure every hand on the forecast: the system against the naive, the adjustment against the system.

Sources: Rob J. Hyndman & George Athanasopoulos, OTexts, 2021; Robert Fildes et al., International Journal of Forecasting 25(1), 2009 (via Abstract and summaries); Robert Fildes & Paul Goodwin, Foresight 8, 2007

Measures as in Hyndman & Athanasopoulos (2021). Fildes et al.: three manufacturers forecasting monthly, one retailer weekly; the authors see optimism in the upward lean.

TOOL OF THE ISSUE

The forecast review card

One card per product family and month. The naive forecast next to the system's, every hand adjustment with its reason, and a look at whether the error leans one way.

01

Series and level

product family, store or region; how many days had no sales

02

Benchmark

naive or seasonal naive: its error over the same period

03

System forecast

error in units (MAE) and against the benchmark (MASE)

04

Hand adjustments

who changed what, up or down, by how much, and why

05

Value of the adjustment

error after the adjustment against the system forecast before it

06

Lean of the error

months above sales against months below; the errors on a control chart

07

Next month

keep, change the method or the data, or stop the adjustments that did not help

Sources: Rob J. Hyndman & George Athanasopoulos, OTexts, 2021; Robert Fildes et al., International Journal of Forecasting 25(1), 2009 (via Abstract and summaries)

A practice proposed by the editors, after Hyndman & Athanasopoulos (2021) and Fildes et al. (2009). For the monthly plan itself, see No. 59.

Open the tool: [Sigma & Control Chart](#)

stivencatalyst.com/tools/sigma-control-chart/

SOURCES AND METHOD

Every figure has a source.

The figures in this issue come from the sources below. The year shows how recent each one is.

M5 accuracy competition: Results, findings, and conclusions

Spyros Makridakis et al., International Journal of Forecasting 38(4), 2022 · via The authors' preprint, 2020

<https://doi.org/10.1016/j.ijforecast.2021.11.013>

The M5 competition: Background, organization, and implementation

Spyros Makridakis et al., International Journal of Forecasting 38(4), 2022 · via Abstract; the intermittency shares from a later paper

<https://www.sciencedirect.com/science/article/pii/S0169207021001187>

The M4 Competition: Results, findings, conclusion and way forward

Spyros Makridakis et al., International Journal of Forecasting 34(4), 2018 · via Abstract in library records

<https://doi.org/10.1016/j.ijforecast.2018.06.001>

Forecasting: Principles and Practice, 3rd edition

Rob J. Hyndman & George Athanasopoulos, OTexts, 2021

<https://otexts.com/fpp3/>

Effective forecasting and judgmental adjustments: an empirical evaluation and strategies for improvement in supply-chain planning

Robert Fildes et al., International Journal of Forecasting 25(1), 2009 · via Abstract and summaries

<https://doi.org/10.1016/j.ijforecast.2008.11.010>

Good and Bad Judgment in Forecasting: Lessons from Four Companies

Robert Fildes & Paul Goodwin, Foresight 8, 2007

https://forecasters.org/wp-content/uploads/Good-and-Bad-Judgment-in-Forecasting_Issue8.pdf

EDITORIAL METHOD

Each figure was checked for its year, its publisher and what exactly it measures. Where the publisher's page could not be opened, the figure was checked against independent summaries and is marked “via”. The editors' interpretation is marked “Our reading”. Figures that could not be confirmed are not in the issue.



MANAGEMENT REVIEW

Management without theatre.

Every issue, one management question, checked against the best research.

ALL ISSUES

stivencatalyst.com/magazine/management-review.html