# The M5 Competition and the Future of Human Expertise in Forecasting

SPYROS MAKRIDAKIS AND EVANGELOS SPILIOTIS

**PREVIEW** *The M5 forecasting competition is the latest and most widely contested since the first M competition in 1979. Numerous articles have been written appraising the structure of the competitions and the value of their results for forecasting methodology and practice. Mike Gilliland's discussion of the prior competition—the M4—in* Foresight*'s Spring 2020 issue as well as Casper Bojer and Jens Peter Meldgaard's preview of the M5 in our Summer 2020 issue offer nontechnical overviews of these most recent forecasting competitions and their potential influence on the ways we forecast. Other background information on the competitions can be found at* ***https://mofc.unic.ac.cy/****.*

*As there was with the M4, there will be a special issue of the* International Journal of Forecasting*—our sister IIF publication—devoted to a comprehensive assessment of the M5. The preprint "The M5 Accuracy Competition: Results, Findings, and Conclusions" provides a detailed discussion of the participants, methods, and results:* ***https://www.researchgate.net/publication/344487258****.*

*Here, Spyros Makridakis, creator and overseer of the M forecasting competitions (and the person for whom they are named), and Evangelos Spiliotis, his closest collaborator, distill that initial report on the M5 to highlight the winning entries and what they say about the value of expertise in forecast modeling.*

## INTRODUCTION

The M5 competition ran from March to June 2020, attracting more than 5,000 teams representing over 100 countries. It differed from the previous four M competitions in several important ways, some of which were suggested by the discussants of the M4:

- It used hierarchical sales data (from Walmart), starting at the item level and aggregating to that of departments, product categories, and stores in three geographical areas of the U.S. Overall, more than 40,000 daily time series had to be forecast.
- Besides the time-series data, it also included explanatory variables such as price, promotions, day of the week, and special events (e.g. the Super Bowl, Valentine's Day, and Orthodox Easter).
- It asked participants to assess uncertainty in their forecasts by presenting prediction intervals at various levels of confidence from 50% to 99%.
- Many of the time series displayed intermittency, a data characteristic common in practice but not explored in the prior M competitions.

Like the M4, the M5 competition included separate tracks for *forecast accuracy* and *uncertainty* to emphasize their distinct significance for the theory and practice of forecasting. The former attracted over 7,000 competitors and close to 90,000 submissions while the uncertainty track drew more than 1,100 competitors and close to 10,000 submissions.

## WINNING METHOD ENTRIES

The top five methods of both challenges have been successfully replicated by the M5 team. Their codes (except for one top method restricted by proprietary rights) have been deposited in GitHub, where they are available for download (***https://github.com/Mcompetitions/M5-methods***).

The accuracy track was won by YeonJun In, a senior undergraduate majoring in environmental and industrial engineering at Kyunghee University in South Korea. Remarkably, YeonJun is a newcomer to forecasting, having completed a course on the theory and applications of machine learning (ML) only three years ago. Since then, he has contributed to data-science projects and participated in Kaggle competitions alongside many Kaggle grandmasters.

He is the recipient of the winning $30,000 in prize money, $25,000 for the top prize and $5,000 as the top student submission. His accomplishment is all the more surprising because the competition included many data scientists with long experience in the field. In addition, third place went to a team of two PhD students, also studying in a South Korean university.

YeonJun's winning approach was the *lightGBM* method, for the *Light Gradient-Boosted Machine* form of ML. Although the daily Walmart sales data used in the M5 may seem like a very specific application, hierarchical data on unit sales describes the context of most retail firms, which must predict daily sales to efficiently operate their value chain and determine appropriate inventory levels. For this reason, the findings of the M5 are of great practical relevance.

For those interested, the data can be downloaded from ***https://forecasters.org/resources/time-series-data/*** , where you will also find the data of the four previous competitions.

### *The Light Gradient-Boosted Machine (LightGBM) ML Method*

YeonJun provided this information describing his winning method and its successful implementation.

LightGBM is a free, open-source, tree-based gradient-boosting ML algorithm that had previously attracted attention and praise by winning several Kaggle data-science retail sales competitions (Bojer and Meldgaard, 2020). It owes its attractiveness not just to its high accuracy, but faster speed and greater efficiency than other gradient-boosting algorithms: it can handle large volumes of data while requiring lower memory requirements than similar algorithms. Technical aspects of lightGBM are beyond the scope of this article but can be found in publications such as Gupta (2019). Nearly all of the top 50 methods of the M5 accuracy competition used some version of lightGBM.

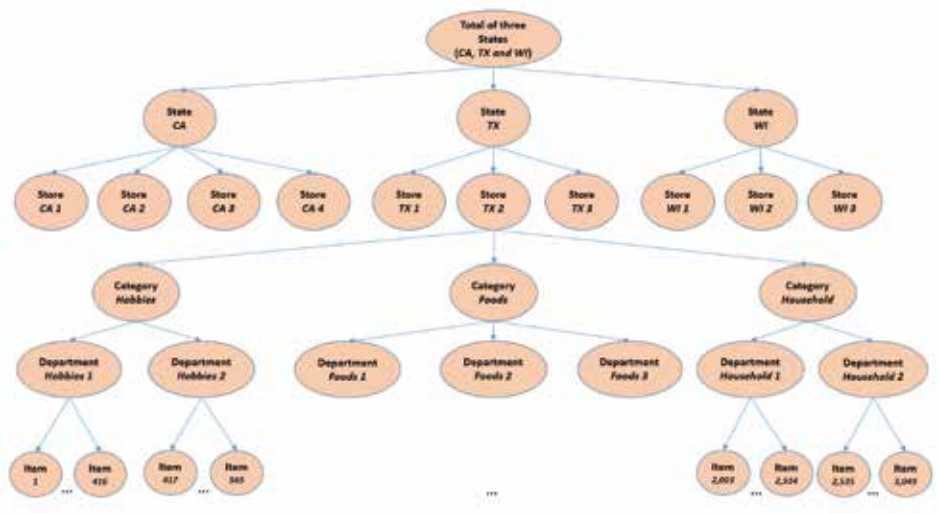
LightGBM is based on GBDT (Gradient-Boosting Decision Trees), an ensemble of decision trees where the errors are trained sequentially from a previously trained tree to the next, improving accuracy at each iteration. One innovation of this method is the use of *an equal weighting to average the forecasts of many complementary models*, each trained by lightGBM at a different hierarchical level. The objective was to produce the most accurate forecasts possible for the product-store level, using data per store (10 models), store-category (30 models), and store-department (70 models) levels. The overall product hierarchy is shown in **Figure 1**.

The models also considered features such as calendar information, special events, promotions, prices, and unit sales data.

## NEW LESSONS

Starting in the early 1970s, forecasting with the popular ARIMA models of that era required considerable statistical

**Figure 1. Product Hierarchy for the M5**




knowledge and expertise: the forecaster had to inspect two or more autocorrelation functions of the data to judgmentally identify an appropriate model specification and apply the Box-Jenkins methodology (Box and Jenkins, 1975). Once a chosen model had been estimated, its residuals had to be inspected to determine if the model was appropriate to be used to forecast beyond the available historical data. There was little concern about overfitting, or that the future course of the time series may diverge from the past history. Still, the Box-Jenkins methodology was considered the forecasting bible, with more than 50,000 citations and four editions of the 1975 text.

Today's software enables ARIMA models to be selected automatically, using information criteria, while also partitioning the historical data to provide a test set for measuring forecast accuracy, a common form of cross-validation. Additionally, we have learned that averaging the forecasts of multiple models frequently improves overall forecasting accuracy in comparison to a selection of a "best" single model.

Other methods of statistical forecasting have also benefitted from automation in model selection. For example, model selection in exponential smoothing was automated using a state space (Hyndman and colleagues, 2002) that proved to produce forecasts that were more accurate than selection based on judgment.

Note too that these methodologies emphasized point forecasts while largely ignoring the inevitable uncertainty surrounding such forecasts.

### *Erroneous Assumptions*

Our forecasting competitions over the decades have rendered erroneous several traditional assumptions, some of which continue to persist:

- A single, "best" model exists for each time series.
- Human judgment is required to decide the method to be used and to select the "best" model to forecast.
- Such models can be estimated using all available data and without concern about overfitting.
- Model forecasts will be as accurate as the model fits to the past data.
- The inevitable uncertainty around the point forecasts can be assumed to be normally distributed.

### *Lessons from the Forecasting Competitions*

The M competitions and other empirical studies have been instrumental in moving the forecasting field away from these erroneous assumptions:

- by showing that ensembles of methods/models are more accurate than single methods;
- by demonstrating the objectivity of systematic forecasting methods and

their superior performance in comparison to those of judgmental predictions, influenced by undue optimism and other human biases;

- by revealing that statistically sophisticated methods achieved a better accuracy in fitting a model to the available data than forecasting from it to post-sample periods (in contrast, simple methods often outperformed sophisticated methods for post-sample predictions, even as they had lower model-fit accuracy);
- by refocusing method selection to consider forecast accuracy on test data not used to train the model; and
- by learning that the normal distribution does not always characterize the uncertainty in the forecasts, which in many cases is fat-tailed, requiring large and asymmetric intervals around the point forecasts.

## THE HUMAN'S ROLE

The M5 competition has not only confirmed the obsolescence of the five traditional assumptions but has provided an additional revelation: the success of ML methods such as lightGBM is challenging the notion of human supremacy in forecast modeling (and other fields), and raising doubts about the value of forecasting knowledge and experience.

The M4 competition winner was an expert data scientist, Slawek Smyl, who devised an innovative hybrid method that combined statistical and ML features. In the M5, the winning entry was based entirely on the application of an ML model—absent statistical-model components—and implemented by a student with a very limited background in forecasting. This distinction between the winners of the M4 and M5 competitions seems to suggest an emerging future of forecasting where ML methods will be capable of doing all the work, confining the role of the data scientist to fine-tuning the recommended model.

But will this be the future of forecasting? What will be the role of academics and research? And how will the practice of forecasting be affected? While the modern forecasting era has evolved slowly but steadily since its inception by Robert Brown in 1959, we now see the forecasting field dominated by ML and DL advances that require few human inputs. The need for experts to analyze data, select appropriate forecasting methods, and assess uncertainty will become minimal as such tasks will be done automatically by the ML/DL method itself with the role of a data scientist limited to tweaking the method for optimal performance.

Does this mean that statistical knowledge, in its present form, will become outdated, statisticians will become data scientists, and that forecasting research will eventually depend on advancements in data science to improve the accuracy and the estimation of uncertainty? As forecasting tasks become automated, forecasters (both academics and practitioners) will need to develop new ways of justifying their roles. Here are some prospects:

- ➡ Work to improve data quality and trustworthiness. For instance, refine measures of demands more precisely than is done through current data on shipments and orders (Gilliland, 2010). For example, when an item is out of stock the lost sales are not recorded or recorded at the wrong time.
- ➡ Develop new opportunities for data analysis. Some participants in the M5 complained that data on inventory levels was not provided. And expand the number of exogenous/explanatory variables to boost model development. Visibility into future promotions during the forecasting period can be very helpful.
- ➡ Given the widespread use of cloud computing, work to share programs and

**While the modern forecasting era has evolved slowly but steadily since its inception by Robert Brown in 1959, we now see the forecasting field dominated by ML and DL advances that require few human inputs.**

data with supply-chain partners so that information flow is seamless. Vendor-managed inventory systems are an example of this goal.

- As uncertainty has not received much attention in either the academic community or among practitioners, concentrate on improving estimation of uncertainty, and more effectively deal with the risks presented and how to face them.
- Given that ML forecasts are "black box," promote understanding of the factors that affect the forecasts and influence uncertainty, thus enabling greater clarity in setting inventory levels, planning promotional campaigns, and achieving control over future actions.
- Work to improve ML/DL forecasts by incorporating inside knowledge and information about likely changes in established patterns/relationships.
- Expand beyond provision and interpretation of forecast numbers to give more attention to policies for achieving competitive advantage.
- In firms that do not currently use systematic forecasting methods and instead rely on ad hoc judgmental predictions, develop necessary data and know how to take advantage—while avoiding loss of competitive position—of the promising technologies emerging from the M5 (Makridakis and colleagues, 2020 a).

**REFERENCES**


Bojer, C.S. & Meldgaard, J.P. (2020). Kaggle Forecasting Competitions: An Overlooked Learning Opportunity, *International Journal of Forecasting*.

Brown, R.G. (1959). *Statistical Forecasting for Inventory Control*, McGraw-Hill, New York.

Box, G.E P. & Jenkins, G. (1975). *Time Series Analysis, Forecasting and Control*, San Francisco, CA: Holden-Day.

Gilliland, M. (2010). Defining "Demand" for Demand Forecasting, *Foresight*, Issue 18, 4-8.

Gupta, R. (2019). LightGBM - Another Gradient Boosting Algorithm, March 28. ***https://rohitgr7.github.io/lightgbm-another-gradient-boosting/***

Hyndman, R.J., Koehler, A.B., Snyder, R.D. & Grose, S. (2002). A State Space Framework for Automatic Forecasting Using Exponential Smoothing Methods, *International Journal of Forecasting*, 18:439–454.

Makridakis, S., Bonnell, E., Clarke, S., Fildes, R., Gilliland, M., Hoover, J. & Tashman, L. (2020a). The Benefits of Systematic Forecasting for Organizations: The UFO Project, *Foresight*, Fall 2020, 45-56.

Makridakis, S., Spiliotis, V. & Assimakopoulos, V. (2020 b). The M5 Accuracy Competition: Results, Findings and Conclusions. ***https://www.researchgate.net/publication/344487258_The_M5_Accuracy_competition_Results_findings_and_conclusions***



**Spyros Makridakis** is the architect of the M-Competitions, a founder of the International Institute of Forecasters, Director of the Institute For the Future (IFF) and the Makridakis Open Forecasting Center (MOFC) at the University of Nicosia and author or coauthor of numerous books and studies that have advanced the practice of forecasting.

**makridakis.s@unic.ac.cy**



**Evangelos Spiliotis** is a Research Fellow at the Forecasting & Strategy Unit, National Technical University of Athens. He is the co-organizer of the M4 and M5 forecasting competitions and a co-author of the papers describing them.

**spiliotis@fsu.gr**