Contents lists available at ScienceDirect

International Journal of Forecasting

journal homepage: www.elsevier.com/locate/ijforecast

Applicability of the M5 to Forecasting at Walmart

Brian Seaman [*], John Bowman

Walmart, 860 W California Ave, Sunnyvale, CA, 94086, USA

ARTICLE INFO

Keywords:
Retail Forecasting
M5 Competition
Practitioner
Walmart
Metrics

ABSTRACT

The M5 Forecasting Competition, the fifth in the series of forecasting competitions organized by Professor Spyros Makridakis and the Makridakis Open Forecasting Center at the University of Nicosia, was an extremely successful event. This competition focused on both the accuracy and uncertainty of forecasts and leveraged actual historical sales data provided by Walmart. This has led to the M5 being a unique competition that closely parallels the difficulties and challenges associated with industrial applications of forecasting. Like its precursor the M4, many interesting ideas came from the results of the M5 competition which will continue to push forecasting in new directions.

In this article we discuss four topics around the practitioners view of the application of the competition and its results to the actual problems we face. First, we examine the data provided and how it relates to common difficulties practitioners must overcome. Secondly, we review the relevance of the accuracy and uncertainty metrics associated with the competition. Third, we discuss the leading solutions and their implications to forecasting at a company like Walmart. We then close with thoughts about a future M6 competition and further enhancements that can be explored.

Introduction

The M5 Forecasting Competition, the fifth in the series of forecasting competitions organized by Makridakis and the Makridakis Open Forecasting Center at the University of Nicosia, focused on both the accuracy and uncertainty of forecasts and leveraged actual historical sales data provided by Walmart (Makridakis et al., 2020a, 2020b). The dual hierarchy of item categories and store locations as well as supplemental data such as prices and events provided researchers and practitioners the opportunity to incorporate a variety of statistical methods and machine learning models. This has led to the M5 being a unique competition that closely parallels the difficulties and challenges associated with industrial applications of forecasting. Similar to its precursor the M4 (Makridakis et al., 2018), many interesting ideas came from the results of the M5 competition which will continue to push forecasting in new directions.

In this article, we discuss four topics around the practitioners view of the application of the competition and its results to the actual problems we face. First, we examine the data provided and how it relates to common difficulties practitioners must overcome. Second, we review the relevance of the accuracy and uncertainty metrics associated with the competition. Third, we discuss the leading solutions and their implications to forecasting at a company such as Walmart. We then close with thoughts about a future M6 competition and further enhancements that can be explored.

Data

Accurate forecasting is extremely important to most businesses and Walmart is no exception. Forecasts are leveraged throughout the company in domains as varied as finance, real estate, supply chain, marketing, and human resources. This breadth of uses highlights the many

[*] Corresponding author.
E-mail addresses: brian.seaman@walmart.com (B. Seaman), john.bowman@walmart.com (J. Bowman).

https://doi.org/10.1016/j.ijforecast.2021.06.002

kinds of data that must be forecasted. Each domain has its unique set of challenges. For instance, inventory forecasting must depend on notoriously unreliable on-hand inventory data, and sales patterns can be greatly altered by localized weather events. These difficulties, however, can often be reduced to a small set of similar problems such as noisy data, missing data, interrelated data, exogenous factors, and anomalous events. One outcome from the M4 competition was the suggestion of the inclusion of several of these traits by Fry and Brundage (2020), and hence, a driver for the M5 competition was to replicate as close as possible these challenges that practitioners face in the real world to compare various forecasting methods.

To better understand the link of the M5 competition with the practitioner's job, it is best to start with the core of any forecasting task which is understanding the underlying data. There were several unique aspects to the data underlying the M5 competition that distinguishes it from prior competitions. This includes the existence of related hierarchical time series, high levels of sparsity and intermittency, and the introduction of additional explanatory variables.

At its simplest, the goal of the M5 Forecasting Competition was to forecast future product sales. For this purpose, Walmart provided over 5 years of historical sales data for over 3000 products. The products spanned 7 departments which were from 3 diverse product categories. The sales data of these products was from 10 stores located in 3 states in the United States. The individual products and locations of the data were anonymized but was otherwise altered. The individual product/store time series as well as all levels of aggregations gives over 42,000 distinct time series that were required to be forecasted.

A rudimentary analysis of sales data highlights a few noticeable features. First, there is a clear weekly and annual seasonality at aggregate levels. This feature becomes less prominent as one moves down the aggregate hierarchies to the product/store level in which case sales volatility is higher and seasonality far less pronounced. Second, many products have zero sales for large portions of the time frame. This can be due to several reasons but is primarily driven by the limited lifecycle of many products that are sold. In retail, products can vary from year to year or season to season as new versions with minor changes or enhancements are released. There are also cases where products are temporarily unavailable due to inventory shortages. The final driver for zero sales is that some items are really long tail items that only occasionally have sales. The zero sales driven by a combination of lack of availability and limited customer interest can lead to interesting problems that need to be solved. The third noticeable feature of the data is that there is a wide variety of sales patterns present in the products. Items can be consistently high selling or low selling with only one sale or less each day and can also show high variability with occasional huge spikes in sales.

The inclusion of explanatory variables was also a key difference between the M5 and the prior Makridakis competitions. When a practitioner understands the business context of the forecasting task, there are usually clear drivers for changes in the time series observations that can be incorporated in the forecasting process. In the retail domain, price is one clear driver for changes in sales, and hence, it was included in the competition data. As part of Walmart's Every Day Low Prices philosophy, prices do not change frequently, but when they do change, it can have an impact on sales. The M5 data also included several prominent events that tend to impact sales. The most consistent and direct is the Supplemental Nutrition Assistance Program, a United States federal food aid benefit run by the Department of Agriculture. The distribution of this aid is tied closely to food sales and an increase in periphery items. A calendar of additional events such as sporting championships and federal holidays were also included as we have attributed sales changes to these events. The rules of the competition also allowed participants to include additional external sets of data, such as weather, as the dates and locations were already known.

Accuracy metrics

To understand the applicability of findings of the M5, it is valuable to examine the evaluation metric used for the competition and how that relates to the forecasts impacts to business decisions. The evaluation metric selected for the accuracy competition was Root Mean Standardized Squared Error (RMSSE):

$$RMSSE = \left[\frac{\frac{1}{h}\sum_{t=n+1}^{h}(y_t - \widetilde{y}_t)^2}{\frac{1}{n-1}\sum_{t=2}^{n}(y_t - y_{t-1})^2} \right]^{\frac{1}{2}}$$

where the training sample is $y_t, t = 1, \ldots, n$ and the forecast is made for periods $n+1$ to $n+h$. RMSSE is a member of a family of relative mean squared error accuracy metrics proposed by Mincer and Zarnowitz (1969), based in part on earlier work by Theil (1966). It has the benefit of automatic scaling by the accuracy of the naive forecast, which leads to the natural interpretation as the improvement (or worsening!) of forecast accuracy by the method being tested relative to the naive method. It also leads to unbiased forecasts, unlike other metrics such as MAE (which drives the forecast towards the median of the relevant distribution) or sMAPE (which, for series with relatively high coefficients of variation, drives the forecast to be biased arbitrarily high[1]). On the other hand, it is not a useful metric from a business perspective, as, for example, at the lowest level of the product – store hierarchy (the individual item – store combination), we tend to use the forecast for inventory control, for which an absolute error metric is needed to help set safety stock, whereas at the highest possible level of aggregation – the companywide earnings statement – the stock market cares about percentage points of error. Seaman (2018), Goodwin (2020) and Koutsandreas et al. (2021) each provide additional discussion on the impact of several different accuracy measurements.

[1] For example, the sMAPE-minimizing estimator of the mean of any Poisson distribution with true mean less than one equals one. This has obvious implications for using sMAPE as a criterion for low demand item forecasting.

There is nothing to prevent the data scientist from using one metric for model evaluation and another for reporting purposes or even using one metric as the objective function when computing the forecasts using a particular algorithm (e.g., minimizing the Tweedie log-likelihood for low-demand items in a Gradient Boosting Machine,) a second for model evaluation, and a third for reporting purposes. In fact, by analogy with the results of the M5 competition itself, which found that different algorithms were preferable at different levels of the product – store hierarchy, it appears that different metrics are preferable for different "levels" of the forecasting process, or for different consumers of the metric (including the forecast algorithms themselves). If one is willing to accept the hopefully minor resultant inconsistencies between what the optimal forecasts are under the various regimes, this would seem to be the way to go, and is more-or-less the approach we have taken at Walmart. (Naturally this applies to metrics for evaluating the performance of quantile and interval forecasts as well.) RMSSE appears to be a good metric for the purposes to which the M5 competition has put it when looked at from this perspective.

The RMSSEs calculated at the lowest level of the item – store hierarchy were aggregated in two steps, first by a weighted average across all item – store combinations with the weights proportional to the dollar sales of the item at the store up to the appropriate node at each level of the hierarchy, and then, the hierarchy-specific RMSSEs were averaged to give the final single score. Kolassa (2020) observes that the best forecast will depend on how "best" is defined, i.e., on the metric. The dollar sales – proportional weighting scheme for aggregating RMSSE across item - stores seems appropriate and is similar to what we do in Walmart. In addition, low volume items often have their supply chain performance driven primarily by the case pack quantity (e.g., 12 or 24 units per box); it hardly matters whether the average forecast error is 150% or 350% if the true mean demand is 0.2 units over the lead-time + review period and the order quantity is fixed at 24. The resultant forecasts are biased toward forecasting higher volume items well, and, qualitatively, that's usually the way it should be.

The equal weighting across the hierarchy leads to forecast accuracy at each level of the hierarchy being equally important to the overall result, and here, we think there is some room for improvement at the price of greater complexity in the competition design. The concerns surrounding equal weighting of hierarchy levels are similar to those of the more general choice of an accuracy metric. Each level of the hierarchy is usually tied to a different use and may drive unrelated business decisions. In general, it is difficult to find a one size fits all approach, and we expand on this idea and suggest modifications in a later section.

Uncertainty metrics

When working with quantiles, the metrics commonly used to evaluate point forecasts are not appropriate. Consequently, Makridakis et al. (2020b) use a scaled Pinball Loss Function, which is a generalization of the loss function implied by the $l-1$ norm that leads to an estimate of the median. For the related problem of interval estimation, a collection of metrics, scaled and unscaled, was reported, giving a more complete picture of interval estimation performance than any single metric could. This latter approach we found helpful in understanding the broader patterns of how the algorithms performed. We hope this continues in future competitions. The scaling of course is only useful when combining item - store metrics. As with the Accuracy competition, we have our doubts about the usefulness of certain aggregations, in this case aggregation across quantiles as well as across levels of the hierarchy. While it would probably be helpful to have an aggregate metric for "lower tail", "center" and "upper tail" estimates, when working at the item – store level, we typically only care about the upper tail, e.g., 75th percentile and above, for inventory control purposes, whereas for higher-level financial reporting analysis of prediction intervals would be more appropriate.

Algorithm selection

We find the diversity of algorithms among the top 50 performers to be encouraging. Clearly there is a great deal of innovative thinking going on in our profession! We use an implementation of histogram-based eXtreme Gradient Boosting Machines (XGBoost) internally for much of our item – store forecasting; longer horizon forecasts are made using a substantially less computationally intensive algorithm, and our algorithm suite also includes hierarchical and multivariate state space models, which perform differentially well on items with different characteristics. We have also made several investigations of various neural net technologies, with disappointing results, although we clearly have some opportunity to learn from the third place and other finishers. The algorithms selected seem to largely match our intuition at a high level about what works and what doesn't, with the one caveat being that at Walmart scale (~500 million item-store combinations in the U.S. alone) the computational burden of heavily multi-model approaches may become too great – although we might then choose to apply such approaches only to the medium and high demand items and for forecast horizons where accuracy counts the most.

The choice of a specific algorithm implementation – LightGBM vs. XGBoost, for example, – is of course important at this stage of algorithm evolution, where how to make "the best" GBM is very much an active area of research. A few years ago, we compared the two and chose XGBoost. Perhaps now, we should revisit that choice. We would not be surprised if GBM technology in the future evolved to the point at which there were several "species" of GBM, each with its own well-understood domain of applicability (not so well-understood at the present time); today, we have Xtreme GBMs with both leaf-wise and level-wise tree growth, Randomized GBMs, GBMs with dropouts such as DART, and others, but algorithm development is occurring at a rapid pace, and it is not clear what the future holds. There are several other algorithms – Random Forests and Neural Nets come readily to mind – in the same state of flux.

Enhancements for an M6 Competition

We see two opportunities for improvement, one in the analysis of the results and the other in the experimental design.

With respect to the analysis, the equal weighting of the results across the different levels of the hierarchy can cause difficulties translating the findings to practical applications. The forecasts at different levels of the hierarchy are used for quite different purposes, and, consequently, the costs of mis-forecasting are not likely to be even approximately the same. For example, at the item – store (lowest) level of the hierarchy, we are primarily concerned with inventory control, usually over the next few weeks (import items may extend that to a few months.) The losses associated with mis-forecasting at this level are well understood, albeit hard to measure at the item level, and consist primarily of stockout costs, inventory holding costs, and wastage. At the product level, though, we are primarily concerned with aiding the supplier's planning process, usually manufacturing or assembly. The relevant horizon is typically considerably longer – in the apparel business, as much as a year – and the costs of significant mis-forecasting can manifest themselves as shortages across many retailers and concomitant high prices. Significant differences can also be found at the other levels of the hierarchy. Because these costs are in many cases impossible to measure with more than order-of-magnitude accuracy, developing appropriate weights for metric averaging would seem to be impossible.

The equal weighting across levels would be unimportant if it were not for the second issue, namely, that the relative performance of even the top algorithms varied significantly across the different levels of the hierarchy. As Makridakis et al. (2020a) conclude: "… depending on the forecasting task and the nature of the data, different forecasting methods should be used to support decisions and optimize forecasting performance at different aggregation levels". Makridakis et al. did a favor by reporting some of the results at different levels of the hierarchy, but, as a practitioner, forecast technologies that perform well at levels other than the ones we work with (e.g., the item – store level) are of considerably less interest than forecast technologies that perform well at the item – store level, and it would be useful, although of course more work, to see more detailed analyses performed at perhaps four to six of the 12 levels of hierarchy considered. Naturally such an analysis would be enhanced by allowing different entries for the different levels of the hierarchy, which in turn might require a greatly expanded data set to allow for many series at each level – but increased analysis time required could be partially compensated for by not analyzing most of the levels in depth.

This logic applies both to the accuracy and uncertainty competitions; forming a quantile estimate of, e.g., the 95th percentile of the U. S. – wide sales of an individual product is a worthwhile task, and should be evaluated on its own, instead of being evaluated jointly with the ability to estimate, e.g., the 50th percentile of the sales of an entire category at a single store.

We also think the experimental design has some room for improvement. Specifically, it appears to us that the 28-day test and validation periods are too short to provide conclusive results. Some evidence of this may be seen in the gap between the validation and test results; only 48.4% of the teams outperformed the Naive benchmark in the Accuracy study, but that number increased to 63.7% if the actual best algorithm from each team was chosen, indicating that the model selection based on the validation period did not correctly identify the best algorithm in a substantial number of cases. (The corresponding numbers for comparison with the ES_bu benchmark were 7.5% and 12.2%.) Qualitatively similar results were found for the Uncertainty study. This is almost certainly due in large part to random differences between the validation and training datasets (unmodeled systematic differences may also play a part.) In our work at the item – store level we prefer to validate over a full year, although some of the reasoning underlying that involves validating holiday and event forecast quality; we consider three months to be pretty much a minimum validation holdout period and would recommend future competitions extend both the validation and test periods to three or more months if possible.

As mentioned previously, different levels of the hierarchy are leveraged for different business decisions which can align with a need for different forecasting horizons. It is tempting to then try to weight the accuracy across several forecast horizons to come to a single metric as was done with the aggregate accuracy metric across hierarchy levels. While this may be necessary for the implementation of a competition it can lead to the same kinds of negative business consequences. To have a competition more closely represent the practitioner's challenges, it may be worth leveraging the same kind of base data but having multiple competitions, i.e., item/store with a three-month horizon and company aggregate with an year-long horizon.

A final note

The role of the statistician/data scientist seems to us to be transitioning from that of a craftworker to those of a machine operator and an engineer. An analogy: before automatic looms were invented, weaving was a craft; there were many weavers producing goods of highly variable quality, and the better weavers were quite expensive to hire. As the early automatic looms became more widespread, the loom operators became less knowledgeable in how weaving was performed as a craft but specialized more in understanding how the automatic looms worked, due mostly to very frequent machine breakdowns and on-the-fly customizations required. Meanwhile, a new profession arose – engineers who kept improving automatic loom technology. (The analogy is not perfect; statisticians have always had an "engineering" component mostly located in universities and government organizations.) As loom technology advanced, the skills required to be a loom operator changed again.

It will be interesting to see what the future holds for the skills required to be an effective data scientist practitioner. The Makridakis competitions are not only extremely valuable in their own right; they form a historical record of how the technology has evolved over the years. We, along with Makridakis et al. believe that "…the value of knowledge and experience will become less important for developing accurate forecasting models which will rely on unstructured, agnostic algorithms and require few human inputs". (Makridakis et al., 2020b, p. 32.) However, the value of research into those unstructured, agnostic algorithms will likely remain high for the foreseeable future.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

Fry, Chris, & Brundage, Michael (2020). The M4 forecasting competition - a practitioner's view. *International Journal of Forecasting*, *36*(1), 156–160.

Goodwin, P. (2020). Performance measurement in the m4 competition: Possible future research. *International Journal of Forecasting*, *36*, 189–190.

Kolassa, S. (2020). Why the "best" point forecast depends on the error or accuracy measure. *International Journal of Forecasting*, *36*(1), 208–211.

Koutsandreas, Diamantis, Spiliotis, Evangelos, Petropoulos, Fotios, & Assimakopoulos, Vassilios (2021). On the selection of forecasting accuracy measures. *Journal of the Operational Research Society*, http://dx.doi.org/10.1080/01605682.2021.1892464.

Makridakis, S., Assimakopoulos, V., Chen, Z., & Spiliotis, E. (2020a). The M5 uncertainty competition: Results, findings, and conclusions. ResearchGate (preprint), https://www.researchgate.net/publication/346493740.

Makridakis, S., Assimakopoulos, V., & Spiliotis, E. (2020b). The M5 accuracy competition: Results, findings, and conclusions. ResearchGate (preprint), https://www.researchgate.net/publication/344487258.

Makridakis, S., Spiliotis, E., & Assimakopoulos, V. (2018). The M4 competition: results, findings, conclusion and way forward. *International Journal of Forecasting*, *34*(4), 802–808.

Mincer, J., & Zarnowitz, V. (1969). The evaluation of economic forecasts. In *Economic forecasts and expectations: analysis of forecasting behavior and performance*. NBER.

Seaman, B. (2018). Considerations of a retail forecasting practitioner. *International Journal of Forecasting*, *34*, 822–829.

Theil, H. (1966). *Applied economic forecasting*. North-Holland.