Kaggle forecasting competitions

An overlooked learning opportunity

Bojer, Casper Solheim; Meldgaard, Jens Peder

Published in:
International Journal of Forecasting

DOI (link to publication from Publisher):
10.1016/j.ijforecast.2020.07.007

Publication date:
2021

Document Version
Accepted author manuscript, peer reviewed version

Citation for published version (APA):
Bojer, C. S., & Meldgaard, J. P. (2021). Kaggle forecasting competitions: An overlooked learning opportunity. *International Journal of Forecasting*, *37*(2), 587-603. https://doi.org/10.1016/j.ijforecast.2020.07.007

Kaggle's Forecasting Competitions: An Overlooked Learning Opportunity[★]

Casper Solheim Bojer[a,*], Jens Peder Meldgaard[a]

[a] *Department of Materials and Production, Aalborg University. Fibigerstræde 16, 9220 Aalborg Øst, Denmark*

Abstract

We review the results of six forecasting competitions on the online data science platform Kaggle that have largely been overlooked by the forecasting community. Contrary to the M competitions, the reviewed competitions feature daily and weekly time series with exogenous variables, business hierarchy information, or both. Furthermore, the Kaggle datasets all exhibit higher entropy than the M3 and M4 competitions and are intermittent.

In our review, we confirm the findings of the M4 competition in that ensemble models using cross-learning tend to outperform local time series models and that gradient boosted decision trees and neural networks are strong forecast methods. Moreover, we present learnings on the use of external information and validation strategies and discuss data characteristics' impact on the choice of statistics or machine learning methods. Based on our learnings, we construct nine ex-ante hypotheses for the outcome of the M5 competition to allow empirical validation of our findings.

Keywords: Time series methods, M competitions, Business forecasting, Forecast accuracy, Machine learning methods, Benchmarking, Time series visualization, Forecasting competition review

[★] This research is funded by the Manufacturing Academy of Denmark (MADE) Digital (Grant No. 6151-0 0 0 06B)

[*] Corresponding author

Email addresses: csb@mp.aau.dk (Casper Solheim Bojer), jenspm@mp.aau.dk (Jens Peder Meldgaard)

1. Introduction

Forecasting is concerned with accurately predicting the future and is a critical input to many planning processes in business, such as financial planning, inventory management, and capacity planning. Considerable interest has been devoted in both industry and academia to the development of methods capable of accurate and reliable forecasting, with many new methods proposed each year. Forecasting competitions, where methods are compared and evaluated empirically on a variety of time series, are widely considered the standard in the forecasting community as they evaluate forecasts constructed ex-ante consistent with the real-life forecast setting (Hyndman, 2020).

Multiple forecast competitions have been held in the forecasting community during the past 50 years, with the M competitions gathering the most attention. The most recent competition, the M4 competition, attempted to take stock of new methods developed in the past 20 years and address criticism regarding the design of previous competitions by including more data frequencies, evaluating prediction intervals, and using statistically robust error measures (Petropoulos and Makridakis, 2020).

Despite the improvements to the competition design, the relevance of the findings of the M4 competition for the business forecasting domain has been the topic of some discussion. Here, practitioners have questioned the representativeness of the competition dataset, which they argue, does not represent many of the forecasting tasks faced by business organizations (Darin and Stellwagen, 2020; Fry and Brundage, 2020). The main critique points concern the underrepresentation of high-frequency series at weekly, daily, and sub-daily levels and the lack of access to valuable information external to the time series, such as exogenous variables and the business hierarchy.

The organizers acknowledged both critique points (Makridakis et al., 2020b), and thus further research on the relative performance of methods for forecasting of higher frequency business time series with access to external information is still required. To facilitate this research, the M Open Forecasting Center (2020) announced the M5 competition, which will be hosted on the online data science platform Kaggle and feature daily time series, including exogenous variables and business hierarchy information. Kaggle is a platform that hosts data science competitions for business problems, recruiting, and academic research purposes. Several forecasting competitions addressing real-life high-frequency business forecasting problems with access to external information have already been completed on Kaggle, but the forecasting community has largely overlooked the results of these.

We believe that these competitions present a learning opportunity for the forecasting community and that they can foreshadow the findings of the M5 competition. To provide an overview of what the forecasting community can learn from Kaggle's forecasting competitions, we:

- identify six forecasting competitions featuring daily or weekly time series with access to external information
- analyze the competition datasets and compare them to the M3 and M4 competitions
- benchmark the Kaggle solutions to ensure they add value beyond simple methods
- review the six competitions and contrast the findings with those of the M4 competition
- provide hypotheses for the findings of the M5 competition

2. Background

The M competitions have been highly influential in the forecasting community, as they focused the attention of the community on the empirical accuracy of methods rather than theoretical properties of models. Additionally, the competitions allowed anyone to participate, enabling contestants with different preferences and skillsets to use their favorite models. The competitions' openness allowed a fairer comparison of methods and tapped into the diverse modeling competencies present in the forecasting community. We refer the reader to the article by Hyndman (2020) for a review of the first three M competitions and focus our attention on the M4 competition, which addressed feedback from the previous competitions (Makridakis et al., 2020c) by:

- including higher frequency data in the form of weekly, daily and hourly data,
- requesting prediction intervals to address forecast uncertainty
- emphasizing reproducibility,
- incorporating many proven methods as benchmarks, and
- increasing the sample size to 100,000 time series to address concerns regarding the statistical significance of the findings.

The time series included in the competition were mainly from the business domain and were restricted to continuous time series, i.e., did not allow intermittence or missing values. More than three full seasonal periods were required at each frequency (Spiliotis et al., 2020), except for the weekly time series where only 80 observations were required (Makridakis et al., 2020c).

The findings of the competition can be divided into four main topics: i) complex vs. simple models, ii) cross-learning, iii) prediction uncertainty, and iv) ensembling, see Makridakis et al. (2020c) for further details. On the topic of complex vs. simple models, the competition found that complex ML methods can outperform simple models often used for time series forecasting, with the top two solutions utilizing a neural network and gradient boosted decision trees (GBDT), respectively. This finding disconfirmed the competition organizers' first hypothesis, as they predicted that the performance of simple methods would be close to the most accurate methods. It is important to note that these methods were adapted to forecasting. Out-of-the-box ML models performed poorly as hypothesized by the organizers (Makridakis et al., 2020a). The competition also demonstrated the benefits of cross-learning, where time series patterns are learned across multiple time series. The top two performers both used models estimated on many time series, which differs from the predominant approach of one model per time series. One of the most surprising findings of the competitions concerned the winner's remarkably accurate estimation of prediction uncertainty. The task of accurately estimating uncertainty is a longstanding challenge in the forecasting field, where most methods underestimate uncertainty (Fildes and Ord, 2007). Finally, the competition once again (Granger and Bates, 1969; Hibon and Makridakis, 2000) confirmed that combinations of forecasting methods, known in ML as ensembling, produced more accurate results than single methods.

A relevant question following the M4 competition concerns the generalizability of the findings and how they might be used to improve forecasting practice. Fildes (2020) argues that forecast competitions establish a pool of empirically proven methods for forecasters to use. He emphasizes that no single method is best for all forecasting tasks and, as a result, recommends that forecasters should select among these methods for their specific use case. This selection could be made by conducting internal forecast competitions or selecting the best methods on a similar subset of the M4. The point that no single method is best for all tasks is also made by Petropoulos et al. (2014), who refer to it as ”Horses for Courses”. They propose that time series features can be used to gain insight into method performance and create a model selection framework based on forecast horizon and five time series features: seasonality, trend, cycle, randomness, and the number of observations. Exploring the issue of ”Horses for Courses” further, Spiliotis et al. (2020) examines whether the M competition datasets are diverse and representative of business forecasting, which is a prerequisite for using the dataset for model selection. They use the feature-based instance space method from Kang et al. (2017) to visualize the datasets in two-dimensional space and conclude that the M4 competition is more diverse than previous competitions. However, they do note that the M4 competition dataset does not contain intermittent time series and might have too few high-frequency series to guide model selection (Spiliotis et al., 2020).

The characteristics of the high-frequency time series included in the M4 competition were also a topic of discussion. The winner at the weekly frequency in the M4 competition ForecastPro noted that many of the weekly time series differed from the ones they typically encounter in business forecasting. They requested that future competitions include more typical business time series, such as demand and sales series at monthly, weekly, daily, and hourly frequencies. They also stressed the importance of access to hierarchy information, as this information can be used to identify the right aggregation level for forecasting (Darin and Stellwagen, 2020). Fry and Brundage (2020) made similar points and requested more high-frequency time series, hierarchy information, and access to potentially relevant exogenous variables. Based on their experience, they suggest that methods using cross-learning, machine learning, and meta-models, provide significant improvements over statistical methods for these types of business forecasting tasks. Besides, they speculate that the lack of access to hierarchy information and exogenous variables cause the poor performance of pure machine learning models. This lack might also explain why the top two methods utilizing ML performed comparatively worse in terms of accuracy at the daily and weekly frequencies, although these frequencies typically have larger sample sizes than at the lower frequencies (Makridakis et al., 2020a).

To address the critique points, the M Open Forecasting Center (2020) announced the M5 competition, which require forecasting of 42,840 daily time series of hierarchical sales data starting at the item level and aggregating to that of departments, product categories, and stores in three US states: California, Texas, and Wisconsin. Besides the time series data, the M5 competition includes explanatory variables such as price, promotions, day of the week, and special events, and the majority of the time series will display intermittency. The competition is split into two parallel tracks using the same dataset, each with a cash prize of $50,000 USD. The first requires 28 days ahead point forecasts with reconcilable aggregates on 12 levels of the business hierarchy, and the second requires 28 days ahead probabilistic forecasts for the median and four prediction intervals (50%, 67%, 95%, and 99%).

The Kaggle platform, which hosts the M5 competition, houses a large community of data scientists from a variety of backgrounds that compete in the competitions and participate in the discussion forums by sharing knowledge and discussing potential strategies. In the business problem-focused competitions, companies provide a dataset for a prediction task and typically offer a cash prize for the top performers.

These competitions typically differ from academic competitions in that they focus on solving a problem rather than learning why and when a particular method works. As an example of this difference in focus, the Kaggle competitions provide real-time feedback on submitted predictions in the form of a publicly available leaderboard, which shows a ranked list of the contestants and their scores. Contestants are allowed to submit multiple predictions, which facilitates learning and results in better predictions (Athanasopoulos and Hyndman, 2011). Kaggle bases the final competition results on private leaderboard performance, which is evaluated on an unseen dataset to prevent overfitting to the leaderboard and allow for an ex-ante assessment.

3. Analysis of Competitions

We initially examined the database of competitions from the online data science platform Kaggle, and only competitions focused on forecasting were kept for further consideration. This resulted in a total of nine forecasting competitions in the history of the platform. We decided to exclude the two earliest competitions, *Tourism Forecasting* and *GEFCOM 2012*, as these were academically hosted competitions with published results and learnings (Athanasopoulos et al., 2011; Hong et al., 2014), which reduced the pool of competitions to the following seven competitions:

- Walmart Store Sales Forecasting [1]
- Rossmann Store Sales [2]
- Walmart Sales in Stormy Weather [3]
- Grupo Bimbo Inventory Demand [4]
- Wikipedia Web Traffic Time Series Forecasting [5]
- Corporación Favorita Grocery Sales Forecasting [6]
- Recruit Restaurant Visitor Forecasting [7]

[1] https://www.kaggle.com/c/walmart-recruiting-store-sales-forecasting

[2] https://www.kaggle.com/c/rossmann-store-sales

[3] https://www.kaggle.com/c/walmart-recruiting-sales-in-stormy-weather

[4] https://www.kaggle.com/c/grupo-bimbo-inventory-demand

[5] https://www.kaggle.com/c/web-traffic-time-series-forecasting

[6] https://www.kaggle.com/c/favorita-grocery-sales-forecasting

[7] https://www.kaggle.com/c/recruit-restaurant-visitor-forecasting

All of these competitions feature high-frequency time series and access to either hierarchy information, exogenous variables, or both. After a more thorough review of the datasets for the competitions, we excluded the *Grupo Bimbo Inventory Demand* competition from further review, as the dataset consisted of a maximum of seven observations per time series at the weekly level with many time series having only one observation, making it unsuitable for time series forecasting.

The six competitions selected for review, along with the characteristics of their forecasting tasks, are summarized in Table 1. Four of the six competitions are from the retail domain, with the remaining two being from the web traffic and restaurant domains. While the retail competitions are from the same domain, they vary in terms of the number of time series and aggregation levels. The *Walmart Store Sales* and the *Rossmann* competitions are both at a very high aggregation level in terms of the business hierarchy with relatively few series and require forecasts for dollar sales by store/department/week and store/day, respectively. The *Corporación Favorita* competition and the *Walmart Stormy Weather* competition are, on the other hand, both at a very disaggregated level with forecasts of unit sales being required by product/store/day, but differ in the number of time series.

The remaining two competitions both feature forecasts at a disaggregate level, as daily forecasts of visits by webpage and restaurants are required, but the domains and the number of time series differ. Due to its' domain, the *Recruit Restaurant* dataset contains data on upcoming reservations, which is expected to contain useful information for the forecasting task. The *Wikipedia* dataset is a more traditional large-scale forecasting task, although it provides access to the business hierarchy through the page URL.

The competitions considered are thus varied in terms of their characteristics but represent a more limited subset of the business forecasting tasks faced by companies than present in the M4 competition. On the other hand, the competitions are more representative of business forecasting tasks with access to exogenous variables and the business hierarchy, which allows us to examine the effects of these two factors.

3.1. Analysis of Datasets

The goal of analyzing the identified Kaggle competition datasets is to position these relative to the M3 and M4 competitions in terms of simple time series characteristics such as entropy, seasonality, and trend. In the analysis, we utilize the methodology developed by Kang et al. (2017) to represent a single time series in two-dimensional space, which allows the analysis of large-scale time series datasets.[8]

[8] Interested readers can find the complete analysis at `https://github.com/cbojer/kaggle-project`

Table 1: Selected Kaggle Competitions and their associated forecasting tasks.

Competition	Time Unit	Forecast Unit	#Observations	#Time Series	Forecast Horizon	Accuracy Measure[1]	Exogeneous Variables	Hierarchy Variables	Point / Interval
Walmart Store Sales (2014)	Weekly	$ Sales by Department	143	3331	1-39	Weighted Mean Absolute Error	Temperature, Markdowns, Fuel Price, CPI, Holidays Unemployment Index	Department ID, Store ID, Store Type, Store Size	Point
Walmart Stormy Weather (2015)	Daily	Unit Sales by Product & Store	851-1011	255	1-7	Root Mean Squared Logarithmic Error	Weather	Weather Station ID, Product ID, Store ID	Point
Rossmann (2015)	Daily	$ Sales by Store	942	1115	1-48	Root Mean Squared Percentage Error	Weather, Closures, Promotions, Holidays, Google Trends, Historical Customer Counts	State, Store ID, Store Type, Assortment Type, Competitor Store Information	Point
Wikipedia (2017)	Daily	Views by Page and Traffic Type	970	~145k	12-42	Symmetric Mean Absolute Percentage Error		Country, Access Agent and Page Name	Point
Corporación Favorita (2018)	Daily	Unit Sales by Product & Store	1684	~210k	1-16	Normalized Weighted Root Mean Squared Logarithmic Error	Holidays, Events, Promotions, Oil Prices	Item ID, Item Family, Item Class, Perishable Flag, Store ID, City, State, Store Type, Store Cluster	Point
Recruit Restaurant (2018)	Daily	Visits by Restaurant	478	821	1-39	Root Mean Squared Logarithmic Error	Reservations, Holidays	Restaurant ID, Genre, Area, Coordinates	Point
M5 Competition Accuracy (2020)	Daily	Unit Sales by Product & Store	1941	30490	1-28	Weighted Root Mean Squared Scaled Error	Holidays, Events, Prices, SNAP[2]	Store ID, Product ID, Department, Category, State	Point
M5 Competition Uncertainty (2020)	Daily	Unit Sales by Product & Store	1941	30490	1-28	Weighted Scaled Pinball Loss	Holidays, Events, Prices, SNAP[2]	Store ID, Product ID, Department, Category, State	Interval

[1] See Appendix A for the mathematical notation of accuracy measures

[2] SNAP refers to Supplemental Nutrition Assistance Program

Data Preprocessing. The Kaggle competition datasets are generally messier than the M competition datasets, with most time series exhibiting intermittence and others having little historical data. Hence, the Kaggle competition datasets require some initial preprocessing to allow extrapolation of the time series instance space. We perform all preprocessing using the R packages data.table (Dowle and Srinivasan, 2019) and base (R Core Team, 2019). The preprocessing can be summarized in five steps:

1. Set NA or Negative values to zero.
2. Remove time series with all zero values.
3. If a test set is available, keep only time series present in both the training and test set.
4. Fill in missing values in irregularly spaced time series[9] with zeroes.
5. Remove leading zeros.

[9] The time series are regular, but some contain missing values due to, e.g., unrecorded sales on store closure, such as weekends or holidays

Competition Representativeness. Due to their ability to provide useful information about the M3 competition data, Kang et al. (2017) proposed a set of features *F1, F2, ..., F6* that enable any time series, of any length, to be summarized as a feature vector **F** = *(F1, F2, F3, F4, F5, F6)*:

1. The *spectral entropy (F1)*, as defined by Goerg (2013), measures "forecastability".
2. The *strength of the trend (F2)* measures the influence of long-term changes in the mean level of the time series.
3. The *strength of seasonality (F3)* measures the influence of seasonal factors.
4. The *seasonal period (F4)* explains the length of periodic patterns.
5. The *first-order autocorrelation (F5)* measures the linear relationship between a time series and the one-step lagged series.
6. The *optimal box-cox transformation parameter (F6)* measures if the variance is approximately constant across the whole series.

To calculate the feature vectors, we use the R package **feasts** (O'Hara-Wild et al., 2019), and subsequently, apply principal components for dimensionality reduction using the prcomp algorithm from the R package **stats** (R Core Team, 2019) to project them all in two-dimensional space to allow for easy visualization with the R package **ggplot2** (Wickham, 2016). A similar method is also used by Spiliotis et al. (2020) in their assessment of the representativeness of the M4 competition. In their study, they use the *seasonal period (F4)* defined by the authors of the M4 competition, namely that yearly, weekly, and daily time series have a seasonal period of one, quarterly time series have a seasonal period of four, monthly time series have a seasonal period of 12, and hourly time series have a seasonal period of 24. As such, seasonality cannot be estimated for weekly and daily series since the estimation algorithm requires at least two full seasonal periods.

To enable estimation of seasonality, we decided to substitute the *seasonal period (F4)* for weekly series with 52 weeks and daily series with seven days. The M4 competition requires 80 observations for the weekly series, which means that seasonality cannot be estimated for some of the series. However, this only concerns around 20% of the time series, and these we revert to the original seasonal period of one and set the estimation of seasonality to zero. For the daily series, we chose a seasonal period of seven days, as the *Restaurant Recruit* competition only features 478 observations, and thereby does not allow for estimation of annual seasonality. Furthermore, just 65% of the daily time series in the M4 competition has more than two years of data.

A side-effect of substituting the seasonal period of the weekly and daily time series is that the *seasonal period (F4)* becomes more influential in the dimensionality reduction, which means that time series with different seasonal periods will become more dispersed since the distance in terms of *seasonal period (F4)* between the series with a period of one and a period of 52 will be higher. Furthermore, we question the value of including the *seasonal period (F4)* as a numerical variable in the first place, as this is essentially a categorical variable (hourly, daily, weekly, etc.) that is known in advance. As such, we believe the impact of the seasonal period is best studied by examining how the feature space for the frequencies differ rather than by including it as a feature. In this research, the goal is to compare the M competitions to Kaggle competitions and not examine how frequency affects time series features. Therefore we have chosen to exclude *seasonal period (F4)* from the dimensionality reduction.

Figure 1 depicts the resulting time series instance space for both the M3, M4, and Kaggle competitions with the density of the data illustrated in each hexbin region with low density in dark grey and high density in blue. The figure highlights the differences between the M and Kaggle competition datasets. For the M competitions, the instance space is most densely populated on the right side, symbolizing strong *trend (F2)* and *ACF1 (F5)*. Conversely, all Kaggle competitions have density peaks further to the left, symbolizing higher degrees of *entropy (F1)*. Common for both the M and Kaggle competitions datasets is that they include time series with varying *seasonality (F3)* and *lambda (F6)*. The most likely reason for the discrepancy in *trend (F2)*, *ACF1 (F5)*, and *entropy (F1)* is that ~ 95% of the time series in the M competitions are low-frequency, i.e., either monthly, quarterly, or yearly, whereas all the reviewed Kaggle competitions are high-frequency, i.e., daily or weekly.

Figure 1 also reveal similarities in the positioning of time series with similar aggregation levels in the business hierarchy. The *Rossmann*, *Recruit Restaurant*, and *Walmart Store Sales* competitions are all at a high aggregation level and have density peaks within the same region of the time series instance space. We see that the highly aggregated time series have lower degrees of *trend (F2)* and *ACF1 (F5)* and higher degrees *entropy (F1)* than the majority of the time series in the M4 competition. The *Corporacíon Favorita*, *Walmart Stormy Weather*, and *Wikipedia* competitions are all at low aggregation levels, but the similarity of their position in the time series instance space is not as apparent. Here, we see that the *Corporacíon Favorita* and *Walmart Stormy Weather* are similar in terms of their density peak areas, and both exhibit relatively high degrees of *spectral entropy*, low degrees of *trend (F2)* and *ACF1 (F5)*, and varying degrees of *seasonality (F3)* and *lambda (F6)*. On the contrary, the *Wikipedia* competition display higher *trend (F2)* and *ACF1 (F5)* than the other competitions with low aggregation levels.

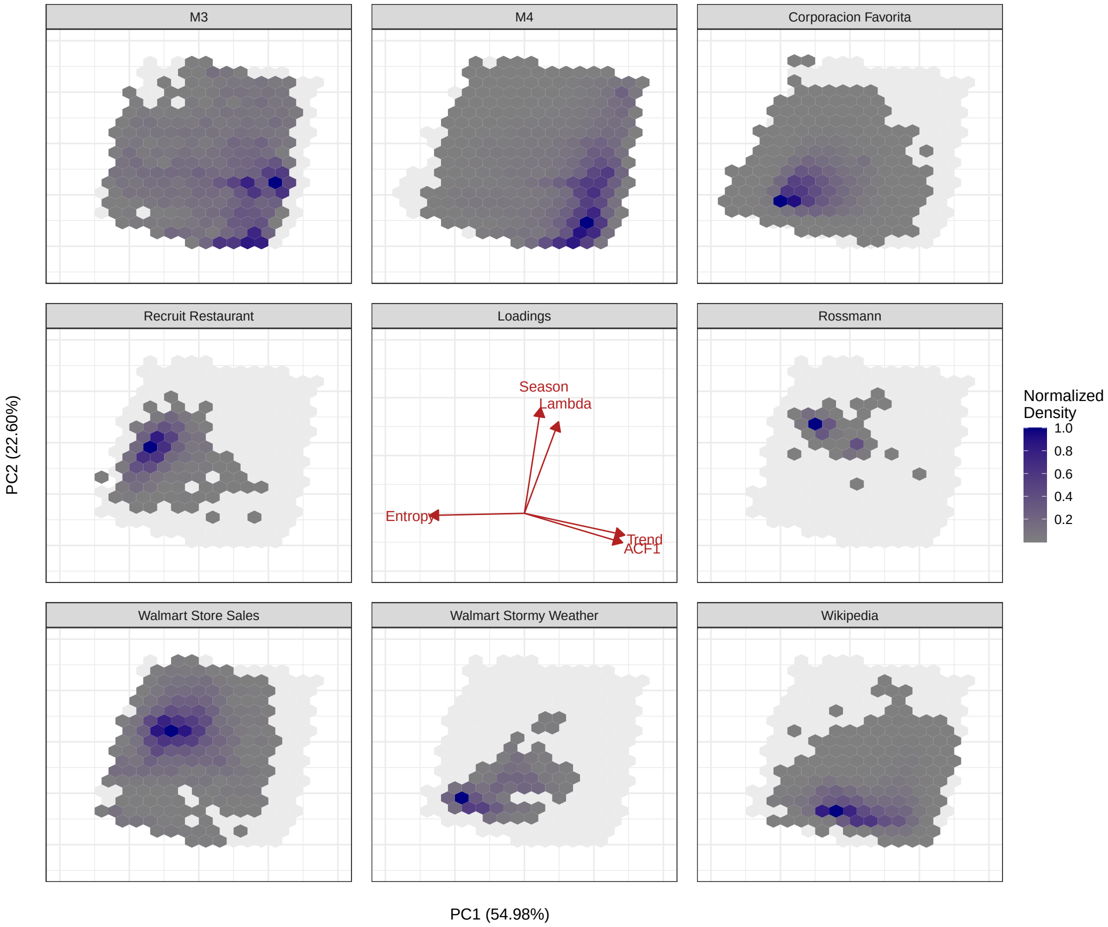

Figure 1: Hexbin plots of the time series instance space of M and Kaggle competitions. The color of each hexbin illustrates the density of time series positioned in that particular field of the instance space, with blue symbolizing high density and dark grey low density. Additionally, the instance space of the M4 competition is illustrated as a light gray background on all plots, except the M4 plot, where the light gray background illustrates the combined instance space of all competition, but the M4 competition. Furthermore, the x- and y-axis titles show the percentage of the total variance, which the 1st and 2nd principal component can explain.

We expect that the dissemblance in *entropy (F1)* between the M competitions and the Kaggle competitions, is to some extent, caused by the selection criteria that prohibit intermittency in the M competitions since, contrary to the M competitions, all of the Kaggle competitions feature some intermittent time series. Explicitly, we find that more than 98% of the time series in the *Corporacı́on Favorita*, *Walmart Stormy Weather*, *Recruit Restaurant*, and *Rossmann* competitions exhibit some degree of intermittency and that approximately 16% and 26% of the time series in the *Walmart Stores Sales* and *Wikipedia* competitions, respectively, exhibit intermittency.

To summarize, we find that the M competition datasets cover a significant part of the time series instance space, but that most of the time series in these are characterized by relatively high *strength of trend (F2)* and *first order autocorrelation (F5)* compared to the those in the reviewed Kaggle competitions. Additionally, we find *spectral entropy (F1)* to be the dividing factor between the M and Kaggle competition, where especially the *Corporación Favorita* competition stands out and even extends the feature instance space beyond that of the M4 competition. As stated, we believe the dissemblance in *spectral entropy (F1)* to be driven by the allowance of intermittent series in the Kaggle competitions. Therefore, we argue that the inclusion of time series with higher degrees of *spectral entropy (F1)* and intermittence would improve the representativeness of future competition datasets.

3.2. Benchmarking Kaggle Solutions

The fundamental premise of our article is that the learnings from top-performing solutions to the Kaggle competitions are valuable. For this to hold, the solutions should as a minimum outperform simple and proven time series forecasting methods. To verify whether this was the case, we benchmarked the solutions against two simple forecasting methods, the naïve and seasonal naïve methods. These methods are often used to identify whether a forecasting method or process adds value in, e.g., forecast value added (FVA) analysis (Gilliland, 2011) or in the MASE forecast accuracy measure (Hyndman and Koehler, 2006). We choose these two methods as they are simple and robust to missing data, which is present in all of the Kaggle competitions. Other often-used benchmarking methods, such as the theta and exponential smoothing models, require a time series without missing data. Therefore, they have not been used, as that would require an imputation procedure to fill in missing values before forecasting.

To construct the benchmark, we produced forecasts for all the competitions using the naïve and seasonal naïve methods. Some of the competitions required forecasts for time series that were not present in the training dataset, and we had to use a fallback method. As a fallback, we used the mean at the next level of the forecast hierarchy to conduct some simple cross-learning. In cases where data was still missing at the next level, we proceeded up the hierarchy until data was present.

We use relative errors to measure the performance of the benchmarks and the top 25 solutions in the Kaggle competitions. To calculate the performance difference, we use the percentage difference relative to the 1st place:

$$\%Difference_{1\text{st},\,\text{n}^{\text{th}}} = \frac{Score_{\text{n}^{\text{th}}} - Score_{1\text{st}}}{Score_{1\text{st}}} * 100$$

Where Score refers to the accuracy measure used in the competition. We use the same accuracy measure as in each of the competitions for two reasons. Firstly, the chosen accuracy measure reflects the business forecasting task, and secondly, the test set is not available, and thus it is not possible to calculate other accuracy measures. A consequence is that the differences are not directly comparable across competitions, as a 20% difference in, e.g., root mean squared logarithmic error is not necessarily the same as a 20% difference in weighted mean absolute error.

Figure 2 shows the benchmarking procedure's results for the better of the two benchmarking methods and the top 25. Across the board, the first place solutions provide improvements above 25% over the simple benchmarks. The performance improvements are particularly striking for the *Rossmann* and *Corporación Favorita* competitions, where the benchmarks are more than 100% worse than the 1st place. It is also evident that some competitions have been much closer than others. In the *Corporación Favorita*, *Recruit Restaurant* and *Walmart Stormy Weather* competitions the gap between the 1st and 25th place is relatively small, meaning that any differences in performance could be due to randomness. On the other hand, the difference is quite significant in the other three competitions, suggesting that performance differences in the top are meaningful. In these three competitions, the difference between the 1st and 2nd place was also more than 2.5%, which suggests that the winner's strategy has an edge over the second placer. Overall, it is clear from the benchmarks that the Kaggle solutions all add value above simple time series benchmarks, and further attention is warranted.

4. Competition Review

To conduct the review, we read through the Kaggle forum posts for each of the competitions. We gathered all information on the solutions posted by contestants, including both textual descriptions and code. Solutions in the top 25 were considered for review in each competition to focus on top performers. Additionally, the forums were examined for the application of simpler methods, such as historical averages or proven forecast methods, to investigate the improvement obtained over the use of simpler methods. Table 2 shows the reported solutions for the top 25 in each competition that we identified during the review process, along with a codification of the methods used. Blank cells indicate contestants that did not describe their methods on the forums. It is evident from the table that a significant part of contestants does not report

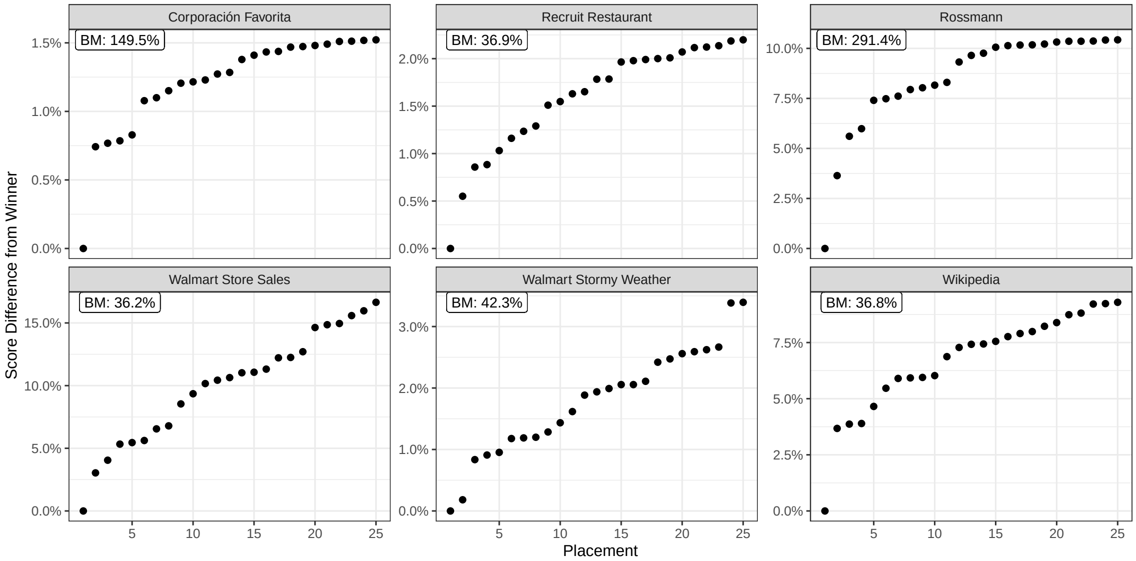

Figure 2: Performance of the top 25 and the best benchmark in each competition. The performance is calculated as the percentage difference in score compared to the winner. The label in the top-left corner of each subplot shows the performance for the best benchmark method.

their methods, although the top performers usually do. This is a limitation of learning from the Kaggle competitions, and we discuss the implications of this in section 5.3. An analysis of table 2 also reveals another pattern: the two Walmart competitions mainly featured time series and statistical models in the top 25, whereas the four later competitions mostly feature GBDT and neural networks. In the following sections, we present detailed findings of the review for each of the six competitions.

4.1. Walmart Store Sales Forecasting (2014)

The Walmart Store Sales forecasting competition is the oldest of the competitions considered. The competition tasked contestants to provide forecasts of Weekly Sales in $ by department and store for a horizon of 1 to 39 weeks. The contestants had access to 33 months of data for the 45 stores and 81 departments, as well as metadata for the stores, holiday information, promotion indicators, weekly temperatures, fuel prices, consumer price index (CPI), and unemployment rate.

Accurate modeling of seasonality and holidays turned out to be crucial in this competition, and top-performing solutions mainly used conventional time series forecasting methods with minor tweaks. The main innovation of the winner was to learn seasonal and holiday patterns globally and use these to denoise the

Table 2: Overview of the reviewed solutions in the top 25 for each of the six Kaggle competitions. Blank white cells indicate that a description of the solution was not available. The text is a codified comma separated list of the methods employed by the solutions. An **(E)** after the method list indicates the use of an ensemble. Blue colored cells indicate the use of cross-learning, and gray cells indicate no use of cross-learning. Codification: **GBDT**: Gradient Boosted Decision Trees, **DTF**: Decision Tree Forests (Random Forest and Extremely Randomized Trees), **TS**: Time series methods (e.g. Exponential Smoothing, ARIMA, Kalman Filter and Moving Averages/Medians), **NN**: Neural networks, **LM**: Linear Regression, **STAT**: Other statistical methods (Polynomial regression, Projection Pursuit Regression, Unobserved Components Model, Principal Components Regression, and Singular Value Decomposition), **ML**: Other ML methods (Support Vector Machines and K-Nearest Neighbors).

Placing	Walmart Store Sales	Walmart Stormy Weather	Rossmann	Wikipedia	Corporación Favorita	Recruit Restaurant
1	STAT, TS (E)	STAT, LM (E)	GBDT (E)	NN (E)	GBDT, NN (E)	GBDT, NN (E)
2	TS, STAT, DTF, ML, LM (E)		GBDT (E)	GBDT, NN, LM (E)	NN (E)	
3	TS	GBDT (E)	NN (E)	NN (E)	GBDT, NN (E)	
4	TS		GBDT (E)	NN	NN (E)	
5	TS	STAT		STAT (E)	GBDT, NN (E)	GBDT, NN (E)
6	TS	LM, DTF, ML, TS (E)		NN (E)	GBDT, NN (E)	
7				NN (E)		GBDT (E)
8	LM			TS	GBDT, NN (E)	GBDT (E)
9	LM (E)					
10	GBDT, LM (E)		GBDT (E)			GBDT (E)
11	TS (E)	GBDT, DTF, ML, LM (E)		NN, TS (E)		GBDT, DTF, TS (E)
12					GBDT, NN (E)	GBDT (E)
13	TS (E)				GBDT, NN (E)	
14				NN, TS (E)		
15					GBDT, NN (E)	
16	GBDT			NN, TS (E)	GBDT, NN (E)	GBDT
17					GBDT	
18						
19		LM		TS (E)		
20						GBDT (E)
21						GBDT, NN, TS (E)
22						
23			GBDT, ML (E)			GBDT, NN (E)
24						
25						GBDT, NN (E)

individual time series. The winner accomplished this using a truncated singular value decomposition (SVD) for each category of time series (department), which was used to reconstruct the individual time series. The effect of the truncation is to remove low-signal variations in the data, effectively filtering out the noise. The denoised time series were then forecasted using local forecasting methods such as STL decomposition in combination with exponential smoothing and ARIMA. Finally, forecasts from these methods were combined with an ensemble of simple forecast models such as seasonal naïve, linear trend and seasonality models, and historical averages. While all the ensembles improved his forecast accuracy, a single model from his ensemble consisting of SVD followed by STL and exponential smoothing would have been accurate enough to win the competition on its own.

An important tweak used by all contestants in the top eight was to adjust the data to lineup holidays from year-to-year, which enabled these to be modeled as part of the seasonal pattern using time series models. ML models did not fare well on their own in the competition and were used mostly as part of an ensemble, also containing time series models. One of the uses came from the second-place solution, which used a combination of ARIMA, unobserved components model, random forest, K-nearest neighbors, linear time series regression, and principal components regression for each department. As such, the models were in the middle on the global vs. local dimension. Interestingly, this relatively complicated ensemble of models did not manage to beat out the simpler first place solution.

The exogenous variables available in the competition, including temperature, fuel prices, consumer price index, unemployment, and information on markdowns, did not prove to be useful in creating accurate forecasts. While some of the top 10 contestants used them, the top two and the 4th place did not, suggesting that they added little value. Other interesting submissions to the competition include the 3rd place due to its simplicity. The submission lined up holidays and used a weighted average of the two closest weeks from last year adjusted for the growth rate of the time series and warm days. This simple solution turned out to be only 4% worse than the best solution. As for standard time series benchmarks, the simple naïve method turned out to be a strong benchmark but was still beaten by more than 20% for all of the top 10 contestants.

4.2. Walmart Sales in Stormy Weather (2015)

The Walmart Sales in Stormy Weather competition featured a slightly different format than the other competitions, as the goal was to forecast the impact of extreme weather on sales. The task of the competition was to provide forecasts of Daily Unit Sales by product and store for a total of 255 time series. The format differed from the other competitions in that forecasts were not required for a future period. Instead, forecasts were required for a ± 3-day window surrounding extreme weather event occurrences, which had been removed from the available data. Thus, the task was not a forecasting task in the purest sense of the word, as observations from after the forecast periods were available. To construct these forecasts, the contestants had access to 28 months of data with some extreme weather events removed for the 44 stores and 111 products, as well as extensive weather information.

The winner of the Walmart Stormy Weather competition used a variation of a common approach in retail forecasting software, which is first to estimate baseline sales and then model the deviations from the baseline using linear regression with exogenous variables. The solution used projection pursuit regression using only time as an input to estimate baseline sales per time series while taking into account trends and potentially yearly seasonality. The deviations from the baseline were modeled with a global L1-regularized linear regression model with interactions using the Vowpal Wabbit library (Vowpal Wabbit, 2007). The main difference from the typical approach in retail forecasting software is thus the use of a more complex smoother than the often-used moving average, as well as a global rather than local regression model. The winner constructed several features[10] from the exogenous variables, including modeling of weekend/weekdays, holidays, and their interactions, along with time information (year, month, day, and trend) and modeling of Black Friday (including lag and lead effects). As expected, the weather data was used in the solution in the form of indicator variables for modeling of threshold effects for precipitation and departure from normal temperatures. However, the winner mentioned in his solution write-up that using weather information did not help forecast performance by much, which is corroborated by other top-placing contestants. This finding is somewhat surprising, as the purpose of the competition was to predict the effect of extreme weather on sales and the actual weather was available instead of a forecast.

Top contestants used several other approaches, such as the use of local Gaussian Process regression by the 5th place solution using mainly date features. Ensembles of various ML models with models such as GBDT using the XGBoost algorithm (Chen and Guestrin, 2016), random forest, SVM, and linear regressions were used successfully by the 3rd, 6th, and 11th place. It is, however, interesting to note that none of these complex ensembles of models, which generally fare well in later Kaggle competitions, did better than the much simpler approach of the winner. This competition also presented the first use of XGBoost, and while placing third, its' performance was not dominating. On the other hand, conventional time series models were not used much in the reported solutions. An exception is the 6th place solution, which used time series models, such as ARIMA, as part of an ensemble. However, ARIMA did not achieve impressive performance on its own, with a performance 17% worse than the winning solution on the public leaderboard.

4.3. Rossmann Store Sales(2015)

The Rossmann Store Sales competition featured the rise of ensembles of global ML models, more specifically the rise of XGBoost. It was also the first competition where a neural network managed to place at the

[10] In this review we use the term exogenous variables to refer to the raw information provided in the competition, and features to refer to the potentially processed inputs to the models.

top three. The competition tasked contestants to forecast Daily Sales in $ by store for a horizon of 1 to 48 days. The contestants had access to 31 months of data for the 1115 stores, as well as metadata for the stores, promotion indicators, holiday information, weather information, and Google Trend statistics.

The winner of the competition outperformed other contestants mainly by adapting the XGBoost model to perform well on time series. This adaptation included the construction of many features using the time series and exogenous variables, as well as a trend adjustment using a ridge regression model to deal with the fact that GBDT cannot extrapolate trends. The main innovations on the feature front consisted of statistics and their rolling versions calculated at different levels of the hierarchy, for different days of the week and promotion periods. Examples include average sales by product, moving averages of sales by product, and average sales by product and promotion status. Additionally, event counters proved useful. These consist of the number of days until, into, and after an event, such as holidays or promotions. The solution also included weather information in the form of precipitation and maximum temperature together with seasonality indicators including Month, Year, Day of Month, Week of Year, and Day of Year to allow for accurate estimation of multiple seasonal effects. A key to good performance with many ML models is the appropriate selection of features and hyperparameters to maximize accuracy without overfitting the training dataset. The strategy used by many contestants was to use hold-out dataset of the same length as the forecast horizon to evaluate the quality of the model and to decide the hyperparameters and select features. The use of ensembling of multiple XGBoost models provided a performance boost of around 5% over the best single model. Variation was introduced into the ensemble by training the model on different data subsets, training models using both direct and iterated predictions, and by including different subsets of features in the models.

Most of the top performers used ensembles of global XGBoost models to create forecasts, but a few of them did include local XGBoost models as part of their ensemble. The features used were generally similar to the winner in that they contain event counters and statistics calculated at various levels of the hierarchy. As such, the exogenous variables related to events, i.e., holidays and promotion, turned out to be essential for obtaining high performance in this competition. This likely explains the significant performance improvements obtained over the seasonal naïve benchmark. The distinguishing feature of the two best solutions is that they used rolling statistics in the form of moving averages or medians as features, thus adapting and utilizing well-known methods in the time series forecasting literature.

The 3$^{\text{rd}}$ place solution successfully used neural networks for the first time in the forecasting competitions on Kaggle. The neural network used was a global fully connected neural network that used the exogenous variables provided in the competition, as well as event counters for holidays and promotions. The time series aspect was handled mainly by the use of seasonality indicators. The seasonality indicators and categorical metadata were modeled using categorical embeddings, where a vector representation of the categories is learned and used by the network for prediction. The solution did not include autoregressive inputs, as is usual for neural networks in forecasting. We refer the reader to the paper posted by the contestants for further details (Guo and Berkhahn, 2016).

The highest scoring simpler method, was the 26th place that used a hybrid approach containing conventional time series models. First, a local ARIMA (with and without exogenous variables) and exponential smoothing models were used to produce forecasts. Afterward, a global XGBoost model was used to forecast their residuals based on weekday, event counters, Google Trends patterns, and weather information to capture the effects of exogenous variables not adequately modeled by the time series models. As such, traditional time series models did not fare well in the competition and were only used together with a global ML model or in the case of moving averages to construct features. The winner managed to beat the time series hybrid model by 11% and a simple benchmark consisting of the median by store, weekday, year, and promotion status by 31%, making it clear that more complex models yielded much better solutions in this competition.

4.4. Wikipedia Web Traffic Forecasting (2017)

The Wikipedia Web Traffic Forecasting competition took scale to another level, requiring forecasts for more than 145.000 time series. It also showcased the power of deep learning for forecasting, which won the competition and took up six places in the top eight. The competition tasked contestants to forecast daily Wikipedia page visits for a horizon of 12 to 42 days. The contestants had access to 32 months of data for the page visits, as well as metadata for the Wikipedia pages.

The winning solution presented both an elegant and accurate deep learning approach without a lot of feature engineering[11], as is typical for solutions using GBDT. The solution consisted of an ensemble of global recurrent neural networks with identical structures. Multiple ensembling approaches were used to reduce the variance of the predictions, as neural network predictions can prove volatile with noisy data. Three models were trained on different random seeds to counteract the randomness of the network's weight initialization. Two approaches were used to prevent sensitivity to the exact number of training iterations used. Firstly, model checkpoints were saved during the training procedure, and averages of the checkpoint

[11] Feature engineering refers to the process of constructing features from exogenous variables or the time series itself.

predictions were used. Secondly, moving averages of neural network weights are used instead of the final weights, also known as stochastic weight averaging (SWA) (Izmailov et al., 2018). The features used in the neural networks consisted of historical page views and categorical variables such as agent, country, site as well as the day of the week. One weakness of recurrent neural networks is that they have difficulties in modeling long-term dependencies, such as is the case with yearly seasonality. The winner found a way around this by including as inputs to the model the page views from a quarter, half-year and year ago. Also, he included the autocorrelation function value at lag 365 and lag 90 to facilitate better modeling of the yearly seasonality. Series were independently scaled to facilitate cross-learning of seasonality and time dynamics, but a measure of scale in the median page views was used to allow the model to learn any potential scale-dependent patterns. A hold-out validation set was used together with an automated hyperparameter tuning algorithm using the Bayesian optimization algorithm SMAC3 (Lindauer et al., 2017) to decide on the hyperparameters of the neural networks. Interestingly, the winner reported that the final performance was relatively insensitive to hyperparameters, with the algorithm finding several models with similar performance.

The other top performers used different neural network architectures, including recurrent neural networks (RNN), convolutional neural networks (CNN), and feedforward neural networks, showcasing that several different architectures can provide similar performance. Likewise, varying degrees of feature engineering was used by the top contestants. The 4th and 6th place solutions used limited feature engineering, while the 2nd place used extensive feature engineering, including using the predictions from the various ensemble models as inputs to another model (referred to as stacking in ML). The takeaway thus seems to be that a multitude of architectures can work and that complex feature engineering is not a requirement for high performing neural network forecasts with this dataset. While neural networks did dominate the competition, another much simpler solution on the 8th place deserves mention. The contestant used a segmented approach, which included Kalman filters to predict high signal series, and a robust approach using the median of moving medians over different windows to predict low signal series. This solution was the only approach in the top that used traditional time series models, and while performing well, it was still around 6% worse than the winning solution.

4.5. Corporación Favorita Grocery Sales Forecasting (2018)

The Corporación Favorita Grocery Sales Forecasting competition is a good demonstration of how the Kaggle community learns from and improves on solutions to previous competitions, as both the gradient boosting approaches used in the *Rossmann* competition and the neural network approaches from the *Wikipedia* competition were utilized heavily by top contestants. The competition tasked contestants to forecast Daily Unit Sales by store and product for a horizon of 1 to 16 days for more than 210.000 time series. The contestants had access to 55 months of data for the 54 stores and 3901 products, as well as metadata for the stores and products, promotion indicators, holiday information, and oil prices.

The winner used a relatively complex ensemble of models consisting of both gradient boosting models and neural network models. One change from earlier competitions was the use of the new and significantly faster gradient boosting library LightGBM (Ke et al., 2017), which makes it easier to experiment with different features and parameters. An innovation in the solution was the training of one model per forecast horizon, rather than one model for all forecast horizons to allow the models to learn what information is useful for each horizon. While yielding good results, this approach does have the trade-off of requiring 16 models rather than 1 model. This approach was used for a LightGBM model as well as a feedforward neural network in an ensemble with two other models. These models consisted of another LightGBM model trained for all horizons and the CNN architecture that placed 6th in the *Wikipedia* competition. The features used in the feedforward neural network and the GBDT models were generally similar to the features utilized successfully in the *Rossmann* competition. The features were mainly rolling statistics grouped by various factors such as store, item, class, and their combinations. The statistics used included measures of centrality and spread, as well as an exponential moving average.

Interestingly, the winner only used very recent data in the models, electing to drop older observations based on validation dataset performance. Thus, the final models used less than a full season of data for model fitting in the form of either one, three, or five months of data, despite multiple seasons being available. Other top placers also favored this approach, such as the 5th and 6th placers. One possible explanation of why this worked despite ignoring the yearly seasonality is the trend present in the data, as well as the short forecast horizon of only 16 days.

No simple approaches were present among the top placers of the competition, which all used similar modeling approaches, consisting of LightGBM paired with feature engineering based on rolling statistics, neural networks inspired by the successful architectures from the Wikipedia competition, or ensembles of the two. The main differences between the solutions were in the details of the feature engineering and architecture, or the validation approach used.

While the hold-out strategy has been used throughout most of the previous competitions to prevent overfitting, several contestants experimented with other validation approaches. One example was the 4th place solution, which held out a certain percentage of time series, thus relying purely on cross-learning for performance estimation. Another interesting validation approach was the use of a combination of grouped K-Fold cross-validation to estimate parameters, and time series cross-validation to estimate model performance. In the grouped K-Fold cross-validation, each time series was restricted to one fold to avoid information leakage across folds, thus also relying purely on cross-learning. The time-series cross-validation used two consecutive hold-out datasets of 16 days to estimate model performance. Despite the interesting aspect of multiple validation approaches working successfully for forecasting, the hold-out approach seems to continue to suffice, as the top three solutions used it.

4.6. Recruit Restaurant Visitor Forecasting (2018)

The Recruit Restaurant Visitor Forecasting competition was a confirmation of the success of previously used methods such as GBDT using rolling statistics, and to some degree of neural networks, in a different domain. The competition tasked contestants to forecast Daily Restaurant Visits by restaurant for a horizon of 1 to 39 days. The contestants had access to 15 months of data for the 821 restaurants, as well as metadata for the restaurants, holiday information, and reservations for restaurant visits made at different times in advance.

The winner in the competition was a team of four contestants, who used an ensemble consisting of the average of their models based on LightGBM, XGBoost, and feedforward neural networks. All of the models used features based on rolling statistics as well as lagged values of restaurant reservations, which presented the main difference from earlier competitions in different domains. Another challenge in the competition was that the test set included the "Golden Week" holiday period, which has significantly different behavior, while contestants only had access to one earlier holiday period in the training dataset. Some contestants discovered an intelligent adjustment to the data to better model these holidays with the little data available by treating holidays as Saturdays and the days prior and preceding as Fridays and Mondays, respectively. This tweak generally gave a significant performance boost when used, as evaluated after the competition by multiple top placers. Using the trick was not necessary to win the competition, as exemplified by the 1st place solution. However, it underlines the value of using domain knowledge and manual adjustments to the data to achieve the best possible performance, similar to the findings of the Walmart Store Sales competition.

The 1st and 5th place solutions used neural networks, but they were not generally as successful as in earlier competitions and were mainly used to add diversity to ensembles. The recurrent and convolutional neural network variants used successfully in the *Wikipedia* and *Corporación Favorita* competitions generally performed slightly worse than models based on boosted decision trees. The 21st, 23rd, and 25th places utilized these methods with around 2% worse accuracy than the 1st place. A potential reason for this could be the size of the dataset, which is smaller than the *Wikipedia* and *Corporación Favorita* datasets by more than a factor 100. Interestingly, a Kalman filter managed to place competitively as in the *Wikipedia* competition with a 33rd place and a performance gap of only 2.4% to the 1st place, highlighting that more traditional time series models can still be viable with exogenous variables available.

As in earlier competitions, most contestants used a hold-out dataset for validation of model performance, although both the 7th and 8th placers surprisingly managed to get high placements using a standard K-Fold validation approach, ignoring the time series nature of the data. Inspired by the innovation from the *Corporación Favorita* competition, a few contestants trained horizon specific models, with one model by the 1st place and the 5th place submission requiring a total of 42 models. A compromise was made by the 11th place contestant, who trained one model per week for six models in total, to still model some of the potential horizon specific effects. However, not all contestants in the top used horizon specific models, suggesting that performance improvements of the approach might not be substantial compared to the growth in the number of models.

5. Discussion

In this section, we pull together the learnings from the six Kaggle competitions and discuss how they contribute to the knowledge base in the forecasting community by:

- summarizing and discussing the findings of our competition review
- discussing the practical applicability of the learnings
- providing nine ex-ante forecasts for the outcome of the M5 competition
- discussing limitations concerned with learning from Kaggle competitions

5.1. Findings

Cross-Learning and Combinations. Our review supports the findings of the M4 competition regarding ensembles vs. single models and cross-learning vs. local models. Ensembles won all of the competitions, and thus this finding continues to hold across different domains and forecasting tasks. Cross-learning was also used by all of the competition winners, although sometimes in combination with local models, which underlines the benefits of cross-learning for time series and motivates further research within this area.

External Information. The performance difference between global and local models in the conducted benchmark suggests that access to the business hierarchy provides cross-learning benefits even higher than those found in the M4 competition. Access to exogenous variables other than the hierarchy provided substantial benefits in some competitions and very small or none in others. Where available, information known in advance such as promotions, holidays, events, and reservations, proved highly useful in most of the competitions. On the other hand, variables that would need to be forecast, namely weather, and macroeconomic variables, did not seem to provide significant benefits, despite the availability of actual values rather than forecasts in the reviewed competitions. A similar finding was obtained in the non-public part of the *Tourism Forecasting* competition (Athanasopoulos et al., 2011), suggesting that this holds in multiple domains.

Statistics vs. Machine Learning. In our review of the six competitions, we did not find one method that dominated all of the competitions. The two earliest competitions, *Walmart Store Sales* and *Walmart Stormy Weather*, were won by innovative use of time series and statistical methods, respectively. The four later competitions were won by non-traditional forecasting methods in the form of either GBDT utilizing rolling and grouped statistics, or neural networks. Additionally, there is a surprisingly similar structure in the top-performing solutions across the four latest competitions. Thus, an interesting question is why the GBDT or neural networks did not perform well in the first two competitions? An apparent reason is that the methods were not mature or even developed at the time of the first two competitions. Neural networks were not used successfully for forecasting in Kaggle competitions before the *Rossmann* competition and their performance in the NN3 competition were unimpressive (Crone et al., 2011). The key to success with neural networks seems to be the use of cross-learning and the adoption of various innovations in terms of architectures, such as LSTM (Hochreiter and Schmidhuber, 1997) and embedding (Guo and Berkhahn, 2016) layers. The first successful GBDT algorithm in the form of XGBoost was not released until after the first Walmart competition. While XGBoost was available and used in the *Walmart Stormy Weather* competition, the method was still new, and the adaptations to the time series domain were not developed. Therefore, it is impossible to say whether the competitions would still be won by time series and statistics methods if held today.

A better question might be: why did time series and statistical methods not perform competitively on the latest four competitions? We believe that the characteristics of the last four competition datasets are better suited to both GBDT and neural networks. The four latest datasets are all characterized by intermittency, and contain external information relevant to the forecasting task in the form of hierarchy information and predictive exogenous variables, such as holidays, events, promotions, and reservations. On the other hand, the *Walmart Store Sales* competition is continuous, has access to hierarchy information, and the exogenous variables contain little useful information, likely due to the high aggregation level. Altogether, this presents ideal conditions for global time series methods. The *Walmart Stormy Weather* competition has the smallest of the competition datasets and has little business hierarchy information, which limits the opportunity for cross-learning. The only exogenous variables provided are related to weather, which did not prove very useful. Furthermore, the availability of data from both before and after the required forecasting periods and the short forecast horizon makes it ideal for statistical smoothing methods, such as the projection pursuit regression employed by the winner. Thus, we find that for disaggregate datasets that are intermittent and contain relevant external information, ML methods outperform both time series and statistical methods, which is in line with the practical experience of forecasters at both Google (Fry and Brundage, 2020) and Amazon (Salinas et al., 2020).

One concern often voiced with regards to more complex methods is their practical applicability and whether the potential accuracy gains justify the added complexity and computational requirements (Gilliland, 2020). The ML methods used in the four most recent Kaggle competitions all require the training of multiple complex models, and they are thus more expensive than popular time series benchmarks in terms of cost and time. In our review, we find that the top solutions generally provide considerable improvements over simple benchmarks, with the seasonal naïve method being between 35% to 290% worse than the winners. As such, the use of more complex methods that effectively use the business hierarchy and exogenous variables should warrant serious consideration for daily and weekly business forecasting tasks, and further research should investigate this trade-off dimension in more detail.

The GEFCOM 2014 and 2017 competitions are from an entirely different domain than the reviewed competition, but also featured high-frequency time series, access to exogenous variables, and competitors using both statistics and ML methods. Looking at the GEFCOM competitions, neither statistics nor ML methods emerge as a clear winner across the competitions, similar to our findings, and consistent with the ”Horses for Courses” hypothesis (Petropoulos et al., 2014). Statistics methods won some of the competitions and performed worse than ML in others, and we speculate that this is due to differences in dataset characteristics of the five competitions. We note that despite these differences, we see methods from both statistics and ML utilized by at least one competitor in the top for all of the GEFCOM competitions. The load forecasting competitions that were won by statistical methods are characterized by relatively strong seasonality and a strong well-understood relationship with one key variable, namely temperature. The wind and solar power forecasting challenges where ML did better are characterized by intermittency, a lower signal-to-noise ratio, and weaker relationships with the exogenous variables. Further research should investigate the relationship between the performance of ML and statistical methods and dataset characteristics in more detail. We suggest to include intermittency and the information content of both exogenous variables and the business hierarchy as influencing factors. Thus, a key question becomes how to operationalize the concept of information content in a manner that generalizes across time series datasets.

Gradient Boosted Decision Trees vs. Neural Networks. As for any differences between GBDT and neural networks, we note that the neural networks outperformed GBDT in the Wikipedia competition, which was very large and contained no useful exogenous variables. In the other three of the latest competitions, both methods placed in the top. Therefore, we speculate that the strength of GBDT is its ability to model external information. Neural networks are the topic of much current forecasting research and were used by the winner of the M4 competition (Smyl, 2020). However, we are not aware of research that uses GBDT in combination with the strategies from the Kaggle competitions. While the second-place solution of the M4 competition used GBDT, it was used as a meta-learner to combine traditional time series forecasting methods (Montero-Manso et al., 2020).

Further research should investigate the use of GBDT for forecasting, given their strong empirical performance in the competitions and several useful properties for forecasting. Since GBDT is based on decision trees, it can learn to deal effectively with in-sample level shifts by partitioning along the time dimension. Additionally, by encoding the business hierarchy using rolling and grouped statistics, it can cross-learn by partitioning on these statistics to pool information from similar time series. Furthermore, the loss function to be optimized is customizable to any function that has well-defined gradients and hessians, e.g., quantile loss as is required to forecast prediction intervals. The main weakness of GBDT is in extrapolating trends. However, Kaggle contestants have developed methods for dealing with this, e.g., ensembling with a linear regression modeling the trend.

Validation Strategies. Throughout all six of the competitions, we find the successful use of a hold-out dataset with length equal to the forecast horizon to validate model performance and prevent overfitting. It is somewhat surprising that we do not see substantial overfitting to the validation set when it is used for multiple evaluations of ML models to select features and hyperparameters. One potential explanation for this is the public leaderboard feedback provided by the Kaggle platform, as a performance drop on the leaderboard would indicate that contestants are likely overfitting the validation set. Therefore, the approach adopted by many contestants in principle corresponds to splitting the data four ways:

1. A training set to estimate models
2. A validation set to evaluate model performance and perform model diagnostics
3. A second small validation set where only the summary performance measure is available to prevent against overfitting
4. The final test set used for evaluating out-of-sample performance.

The second smaller validation set (3) in the form of the public leaderboard is a feature of the Kaggle competition format, in that it facilitates learning and helps avoid overfitting. However, the fact that Kaggle does not make the public leaderboard data available for retraining of models enforces a hold out of the most recent data for validation. The effect of this is essentially that the model has to forecast further than under alternative validation strategies. The M5 competition has addressed this issue by providing contestants with access to the public leaderboard data before forecasts for the final test set are required. Further research should evaluate how the inclusion of a leaderboard, which is later revealed, compares to other established forecast validation strategies such as time-series cross-validation in terms of preventing overfitting.

5.2. M5 Hypotheses

The upcoming M5 competition features a hierarchical dataset from the retail domain generously supplied by Walmart. The competition will require forecasts for more than 40.000 daily time series at the store and product level, and provide contestants with information on prices, promotions, events, and product hierarchy (M Open Forecasting Center, 2020). As such, the forecasting task is very similar to that of the *Corporación Favorita* competition. The main difference from the reviewed competition is the evaluation of prediction uncertainty in addition to prediction accuracy. Based on the learnings from our review, we provide the following ex-ante hypotheses:

- The instance space representation of the time series in the M5 competition will resemble that of the *Corporación Favorita* competition, meaning that entropy is higher and trend and first-order autocorrelation is lower than time series in previous M competitions.

- The winning method will utilize cross-learning, and global and hybrid models will dominate local models.

- Access to hierarchy information will increase the performance gap between local models and models using cross-learning compared to the M4 competition.

- GBDT using feature engineering based on, e.g., rolling statistics and neural networks will both perform well in the competition and outperform existing time series benchmarks in terms of both accuracy and uncertainty.

- To provide prediction intervals, GBDT and neural networks will be adapted by using custom loss functions such as quantile loss, or by adapting the training procedure/architecture to output distributions, which has been the topic of much recent research (see, e.g., Duan et al. (2019) for GBDT and Salinas et al. (2020) for neural networks).

- Ensembles of methods will continue to take up the top slots, as is consistent with the findings from all Kaggle and M competitions. We expect these ensembles to contain both neural networks and GBDT, potentially in combination with other methods.

- Hold-out datasets or time series cross-validation will be used by top placers to avoid overfitting.

- Using known in advance exogenous variables such as prices, promotions, holidays, and other events will provide improvements to forecast accuracy, in line with previous retail research (Fildes et al., 2019) and our review of the Kaggle competitions.

- Contestants will develop innovative strategies to tackle the challenge of hierarchical forecasting, and we expect new neural network architectures and GBDT strategies to utilize this information optimally.

5.3. Limitations

The focus on providing solutions to real-life forecasting tasks in the reviewed competitions has a downside: it provides some limitations to what researchers can infer from the competitions. Lack of access to the test set after the competition has ended means it is impossible to test for significant differences between the performances of solutions or to evaluate performance using alternative error measures. Furthermore, it is not possible to analyze the performance of different solutions on various subsets of the dataset to improve our understanding of the strengths and weaknesses of various methods. Future Kaggle competitions should address this by making the test set available after the competition has ended as in the M4 and M5 competitions.

A major weakness in terms of the practical applicability of the reviewed Kaggle competitions studied is that they do not address prediction uncertainty. Forecasts are always wrong, and thus an estimate of the uncertainty associated with the forecast is key for decision making based on forecasts, e.g., in hedging, capacity planning, and inventory management. As the competitions are based on real-life forecasting tasks, it is surprising that the competition case companies did not demand prediction intervals as part of the forecasting task. An explanation could be that the companies do not currently utilize prediction uncertainty in their planning processes. However, prediction intervals are a requirement for the upcoming M5 competition conducted in collaboration with Walmart, which will hopefully set the standard for Kaggle competitions to come.

Although Kaggle encourages the sharing of solutions, contestants are not required to share their solution or code publicly, which makes learning from the competitions harder and does not enable reproducibility of the results. An ideal solution would be for Kaggle to require contestants to submit their code, enabling reproducibility of results and a full mapping and analysis of the solutions. A less strict alternative could be to make contestants fill out a small survey with questions around the methods and approach used when submitting their final predictions. While this does not address the issue of reproducibility, it would facilitate learning from competitions by enabling the mapping of the solutions, although it will necessarily be less detailed than if code was available.

The lack of publicly-shared solutions also has implications for the validity of our review. It is a possibility that the use of methods such as linear regression or local time series models in the non-reported solutions in the top 25 would change our results. However, we find it highly unlikely that local time series would have been able to perform competitively in the four latest competitions, which we base on the intermittency and influence of exogenous variables in the datasets. The results of our benchmarking also support this. We also find the presence of a systematic reporting bias caused by differences in willingness to share for different solution methods unlikely. Despite these weaknesses, we still believe much can be learned by focusing on the patterns of what worked across the competitions and relating the findings to dataset characteristics. Further research should subject our hypotheses to testing on a variety of datasets, and the upcoming M5 competition will surely serve as a great initial testing ground.

6. Conclusions

Based on our analysis and review of the six recent Kaggle forecasting competitions, we believe that the forecast community has a lot to learn from the Kaggle community in terms of forecasting daily and weekly business time series. In our analysis, we find that the M4 competition dataset contains time series similar to those of the Kaggle competitions, although time series with these characteristics are underrepresented in the M4 competition dataset. Furthermore, the Kaggle datasets differ from the M4 competition in that they provide access to external information, e.g., exogenous variables or business hierarchy, which led to significant improvements in forecast accuracy.

Similar to the findings from the M4 competition, we find that global ensemble models outperform local single models. In contrast to the M4 and the two earlier Kaggle competitions, conventional time series and statistical methods were significantly outperformed by machine learning methods in the four latest Kaggle competitions. We believe that this can be attributed to the machine learning methods' utilization of external information to cross-learn and model the effect of exogenous factors. Additionally, we find a similarity between top solutions in the Kaggle competitions and the top two solutions in the M4 competition, which relied on either gradient boosted decision trees or neural networks. However, to obtain the performance benefits from machine learning methods, several adaptations to the machine learning methods and their validation strategies must be adopted.

We strongly encourage the forecast community to learn from the presented machine learning strategies for time series forecasting and to participate in further development. The M5 competition presents an ideal opportunity for this, as the forecasting task and dataset bear high similarity to some of the Kaggle competitions reviewed in this paper. Therefore, we believe that the learnings from the Kaggle competitions discussed in this paper will foreshadow the results of the M5 competition.

References

Athanasopoulos, G., Hyndman, R.J., 2011. The value of feedback in forecasting competitions. International Journal of Forecasting 27, 845–849. URL: http://www.sciencedirect.com/science/article/pii/S0169207011000495, doi:10.1016/j.ijforecast.2011.03.002.

Athanasopoulos, G., Hyndman, R.J., Song, H., Wu, D.C., 2011. The tourism forecasting competition. International Journal of Forecasting 27, 822–844. doi:10.1016/j.ijforecast.2010.04.009.

Chen, T., Guestrin, C., 2016. XGBoost: A Scalable Tree Boosting System, in: Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Association for Computing Machinery. pp. 785–794. URL: https://doi.org/10.1145/2939672.2939785, doi:10.1145/2939672.2939785.

Crone, S.F., Hibon, M., Nikolopoulos, K., 2011. Advances in forecasting with neural networks? Empirical evidence from the NN3 competition on time series prediction. International Journal of Forecasting 27, 635–660. doi:10.1016/j.ijforecast.2011.04.001.

Darin, S.G., Stellwagen, E., 2020. Forecasting the M4 competition weekly data: Forecast Pro's winning approach. International Journal of Forecasting 36, 135–141. URL: http://www.sciencedirect.com/science/article/pii/S0169207019301177, doi:10.1016/j.ijforecast.2019.03.018.

Dowle, M., Srinivasan, A., 2019. data.table: Extension of 'data.frame'. URL: https://CRAN.R-project.org/package=data.table.

Duan, T., Avati, A., Ding, D.Y., Thai, K.K., Basu, S., Ng, A.Y., Schuler, A., 2019. NGBoost: Natural Gradient Boosting for Probabilistic Prediction. URL: http://arxiv.org/abs/1910.03225.

Fildes, R., 2020. Learning from forecasting competitions. International Journal of Forecasting 36, 186–188. doi:10.1016/j.ijforecast.2019.04.012.

Fildes, R., Ma, S., Kolassa, S., 2019. Retail forecasting: Research and practice. International Journal of Forecasting URL: https://www.sciencedirect.com/science/article/pii/S016920701930192X, doi:10.1016/j.ijforecast.2019.06.004.

Fildes, R., Ord, K., 2007. Forecasting Competitions: Their Role in Improving Forecasting Practice and Research, in: Clements, M.P., Hendry, D.F. (Eds.), A Companion to Economic Forecasting. John Wiley & Sons, Ltd. chapter 15, pp. 322–353. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/9780470996430.ch15.

Fry, C., Brundage, M., 2020. The M4 Forecasting Competition-A Practitioner's View. International Journal of Forecasting 36, 156–160. URL: https://www.sciencedirect.com/science/article/pii/S0169207019301189?via%3Dihub.

Gilliland, M., 2011. Value Added Analysis: Business forecasting effectiveness. Analytics Magazine .

Gilliland, M., 2020. The value added by machine learning approaches in forecasting. International Journal of Forecasting 36, 161–166. URL: http://www.sciencedirect.com/science/article/pii/S0169207019301165, doi:10.1016/j.ijforecast.2019.04.016.

Goerg, G., 2013. Forecastable component analysis, in: International Conference on Machine Learning, pp. 64–72.

Granger, C.W.J., Bates, J.M., 1969. The Combination of Forecasts. Journal of the Operational Research Society 20, 451–468. doi:10.1017/cbo9780511753961.021.

Guo, C., Berkhahn, F., 2016. Entity Embeddings of Categorical Variables.

Hibon, M., Makridakis, S., 2000. The M3-Competition: results, conclusions and implications. International Journal of Forecasting 16, 451–476.

Hochreiter, S., Schmidhuber, J., 1997. Long Short-Term Memory. Neural Computation 9, 1735–1780. doi:10.1162/neco.1997.9.8.1735.

Hong, T., Pinson, P., Fan, S., 2014. Global energy forecasting competition 2012. doi:10.1016/j.ijforecast.2013.07.001.

Hyndman, R.J., 2020. A brief history of forecasting competitions. International Journal of Forecasting 36, 7–14. URL: http://www.sciencedirect.com/science/article/pii/S016920701930086X, doi:10.1016/j.ijforecast.2019.03.015.

Hyndman, R.J., Koehler, A.B., 2006. Another look at measures of forecast accuracy. International Journal of Forecasting 22, 679–688. URL: http://www.sciencedirect.com/science/article/pii/S0169207006000239, doi:10.1016/j.ijforecast.2006.03.001.

Izmailov, P., Podoprikhin, D., Garipov, T., Vetrov, D., Wilson, A.G., 2018. Averaging Weights Leads to Wider Optima and Better Generalization. URL: http://arxiv.org/abs/1803.05407.

Kang, Y., Hyndman, R.J., Smith-Miles, K., 2017. Visualising forecasting algorithm performance using time series instance spaces. International Journal of Forecasting 33, 345–358.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., Liu, T.Y., 2017. Lightgbm: A highly efficient gradient boosting decision tree, in: Proceedings of the 31st International Conference on Neural Information Processing Systems, Curran Associates Inc.. pp. 3146–3154.

Lindauer, M., Eggensperger, K., Feurer, M., Falkner, S., Biedenkapp, A., Hutter, F., 2017. SMAC v3: Algorithm Configuration in Python. URL: https://github.com/automl/SMAC3.

M Open Forecasting Center, 2020. The M5 Competition. URL: https://mofc.unic.ac.cy/m5-competition/.

Makridakis, S., Spiliotis, E., Assimakopoulos, V., 2020a. Predicting/hypothesizing the findings of the M4 Competition. doi:10.1016/j.ijforecast.2019.02.012.

Makridakis, S., Spiliotis, E., Assimakopoulos, V., 2020b. Responses to discussions and commentaries. International Journal of Forecasting 36, 217–223. URL: http://www.sciencedirect.com/science/article/pii/S0169207019300871, doi:10.1016/j.ijforecast.2019.05.002.

Makridakis, S., Spiliotis, E., Assimakopoulos, V., 2020c. The M4 Competition: 100,000 time series and 61 forecasting methods. International Journal of Forecasting 36, 54–74. URL: http://www.sciencedirect.com/science/article/pii/S0169207019301128, doi:10.1016/j.ijforecast.2019.04.014.

Montero-Manso, P., Athanasopoulos, G., Hyndman, R.J., Talagala, T.S., 2020. FFORMA: Feature-based forecast model averaging. International Journal of Forecasting 36, 86–92. URL: http://www.sciencedirect.com/science/article/pii/S0169207019300895, doi:10.1016/j.ijforecast.2019.02.011.

O'Hara-Wild, M., Hyndman, R., Wang, E., 2019. feasts: Feature Extraction And Statistics for Time Series. URL: https://cran.r-project.org/package=feasts.

Petropoulos, F., Makridakis, S., 2020. The M4 competition: Bigger. Stronger. Better. International Journal of Forecasting 36, 3–6. URL: https://www.sciencedirect.com/science/article/pii/S0169207019301116, doi:10.1016/j.ijforecast.2019.05.005.

Petropoulos, F., Makridakis, S., Assimakopoulos, V., Nikolopoulos, K., 2014. 'Horses for Courses' in demand forecasting. European Journal of Operational Research 237, 152–163. URL: https://www.sciencedirect.com/science/article/pii/S0377221714001714, doi:10.1016/j.ejor.2014.02.036.

R Core Team, 2019. R: A Language and Environment for Statistical Computing. URL: https://www.r-project.org/.

Salinas, D., Flunkert, V., Gasthaus, J., Januschowski, T., 2020. DeepAR: Probabilistic forecasting with autoregressive recurrent networks. International Journal of Forecasting 36, 1181–1191. URL: http://www.sciencedirect.com/science/article/pii/S0169207019301888, doi:10.1016/j.ijforecast.2019.07.001.

Smyl, S., 2020. A hybrid method of exponential smoothing and recurrent neural networks for time series forecasting. International Journal of Forecasting 36, 75–85. URL: http://www.sciencedirect.com/science/article/pii/S0169207019301153, doi:10.1016/j.ijforecast.2019.03.017.

Spiliotis, E., Kouloumos, A., Assimakopoulos, V., Makridakis, S., 2020. Are forecasting competitions data representative of the reality? International Journal of Forecasting 36, 37–53.

Vowpal Wabbit, 2007. Vowpal Wabbit. URL: https://github.com/VowpalWabbit/vowpal_wabbit.

Wickham, H., 2016. ggplot2: Elegant Graphics for Data Analysis. URL: https://ggplot2.tidyverse.org.

Appendix A. Competition Accuracy Measures

For the accuracy measures, h refers to the forecast horizon, y_i refers to the actual values, $\hat{y}_i$ to the forecasted values, and w_i to the weight assigned to observation i. The weight, w_i, is used to penalize forecast errors for particular observations or time series, namely perishables in *Corporación Favorita* and promotion periods in *Walmart Store Sales*.

Walmart Store Sales : Weighted Mean Absolute Error (WMAE)

$$= \frac{1}{\sum_{h}^{i=1} w_i} \sum_{i=1}^{h} w_i|y_i - \hat{y}_i|$$

Walmart Stormy Weather : Root Mean Squared Logarithmic Error (RMSLE)

$$= \sqrt{\frac{1}{h}\sum_{i=1}^{h}(log(\hat{y}_i + 1) - log(y_i + 1))^2}$$

Rossmann Store Sales : Root Mean Squared Logarithmic Error (RMSPE)

$$= \sqrt{\frac{1}{h}\sum_{i=1}^{h}(\frac{y_i - \hat{y}_i}{y_i})^2}$$

Wikipedia Web Traffic : Symmetric Mean Absolute Percentage Error (SMAPE)

$$= \frac{1}{h}\sum_{i=1}^{h}\frac{|\hat{y}_i - y_i|}{(|y_i| + |\hat{y}_i|)/2}$$

Corporación Favorita : Normalized Weighted Root Mean Squared Logarithmic Error (NWRMSLE)

$$= \sqrt{\frac{\sum_{i=1}^{h} w_i(log(\hat{y}_i + 1) - log(y_i + 1))^2}{\sum_{i=1}^{h} w_i}}$$

Recruit Restaurant : Root Mean Squared Logarithmic Error (RMSLE)

$$= \sqrt{\frac{1}{h}\sum_{i=1}^{h}(log(\hat{y}_i + 1) - log(y_i + 1))^2}$$