ARTICLE IN PRESS

International Journal of Forecasting xxx (xxxx) xxx

Contents lists available at ScienceDirect

International Journal of Forecasting

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

Kaggle forecasting competitions: An overlooked learning opportunity☆

Casper Solheim Bojer*, Jens Peder Meldgaard

Department of Materials and Production, Aalborg University, Fibigerstrde 16, 9220 Aalborg, Denmark

ARTICLE INFO

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

ABSTRACT

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

In this review, we confirm the conclusion of the M4 competition 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 insights regarding the use of external information and validation strategies, and discuss the impacts of data characteristics on the choice of statistics or machine learning methods. Based on these insights, we construct nine ex-ante hypotheses for the outcome of the M5 competition to allow empirical validation of our findings.

1. Introduction

Forecasting is concerned with accurately predicting the future and it provides critical inputs for many planning processes in business, such as financial planning, inventory management, and capacity planning. There has been considerable interest in both industry and academia in the development of methods that are capable of accurate and reliable forecasting, and many new methods have been proposed each year. In forecasting competitions, methods are compared and evaluated empirically based on a variety of time series, and they are widely considered the standard by the forecasting community because they evaluate forecasts constructed ex-ante and they are consistent with real-life forecasting settings (Hyndman, 2020).

Many forecasting competitions have been held by the forecasting community during the past 50 years, but the M competitions have attracted the most attention. The most recent was the M4 competition, which attempted to test new methods developed in the past 20 years and it also addressed criticisms regarding the design of previous competitions by including more data frequencies, as well as evaluating prediction intervals and using error measures with better statistical properties (Petropoulos & Makridakis, 2020).

Despite these improvements to the competition's design, the relevance of the findings obtained from the M4 competition for the business forecasting domain have been subject to criticism. In particular, practitioners have questioned the representativeness of the data set used in the competition, which they argued is not representative of many of the forecasting tasks conducted by business organizations (Darin & Stellwagen, 2020; Fry & Brundage, 2020). The main criticisms concern the underrepresentation of high-frequency series at weekly, daily, and sub-daily levels, and the lack of access to valuable information

☆ This research was funded by the Manufacturing Academy of Denmark (MADE) Digital (Grant No. 6151-0 0 0 06B).

* Corresponding author.
E-mail addresses: csb@mp.aau.dk (C.S. Bojer), jenspm@mp.aau.dk (J.P. Meldgaard).

https://doi.org/10.1016/j.ijforecast.2020.07.007
0169-2070/

external[1] to the time series, such as exogenous variables and the business hierarchy.

The organizers of the M4 competition have acknowledged both criticisms (Makridakis et al., 2020b), and thus further research is still required regarding the relative performance of methods for forecasting higher frequency business time series with access to external information. To facilitate this research, the M5 competition was announced by M Open Forecasting Center (2020), which will be hosted by the online data science platform Kaggle and feature daily time series, as well as including exogenous variables and business hierarchy information. Kaggle is a platform that hosts data science competitions for business problems, recruitment, and academic research purposes. Several forecasting competitions that address 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 obtained.

We consider that these competitions present a learning opportunity for the forecasting community and that they may foretell the findings of the M5 competition. In the following, to provide an overview of what the forecasting community might learn from forecasting competitions on Kaggle, we:

- Identify six forecasting competitions featuring daily or weekly time series with access to external information;
- Analyze the competition data sets and compare them with those used in the M3 and M4 competitions;
- Benchmark the Kaggle solutions to ensure that they add value beyond simple methods;
- Review the six competitions and contrast their findings with those obtained in the M4 competition;
- Provide hypotheses regarding the findings of the M5 competition.

2. Background

The M competitions have been highly influential in the forecasting community because they focused the attention of the community on the empirical accuracy of methods rather than the theoretical properties of models. In addition, the competitions allowed anyone to participate, thereby enabling contestants with different preferences and skill sets to use their favorite models. The openness of the competitions facilitated fairer comparisons of methods and tapped into the diverse modeling competencies present in the forecasting community. The study by Hyndman (2020) reviewed the first three M competitions so we 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 forms 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 competitions came mainly from the business domain and they were restricted to continuous time series, i.e., they 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 obtained in 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). In terms of complex vs. simple models, the competition found that complex machine learning (ML) methods could outperform the simple models that are often used for time series forecasting, where the top two solutions were obtained by utilizing neural networks and gradient boosted decision trees (GBDT). These findings failed to support the first hypothesis proposed by the competition's organizers who predicted that the performance of simple methods would be similar to that of the most accurate methods. It is important to note that these methods were adapted to forecasting, whereas 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 based on many time series, which differs from the dominant approach that uses one model per time series. One of the most surprising findings obtained from the competitions concerned the remarkably accurate estimation of the prediction uncertainty by the winner. The accurate estimation of uncertainty has been a major challenge in the forecasting field for many years because most methods underestimate uncertainty (Fildes & Ord, 2007). Finally, the competitions again (Granger & Bates, 1969; Hibon & Makridakis, 2000) confirmed that combinations of forecasting methods, known as ensembling in ML, produced more accurate results than single methods.

An important question raised by the M4 competition concerns the generalizability of the findings and how they might be used to improve forecasting practice. In Fildes (2020), it was argued that forecasting competitions can establish a pool of empirically proven methods for use by forecasters. It was also emphasized that no single method is best for all forecasting tasks, and thus it is recommended that forecasters select an appropriate method for their specific use case, such as by conducting internal forecasting competitions (Fildes, 2020) or selecting the best methods based on a similar subset from the M4 competition (Makridakis et al., 2020b). The conclusion that no single method is best for all tasks was also made by Petropoulos et al. (2014), who referred to it as "horses for courses". They proposed that time series features can be used to obtain insights into the

[1] External information refers to both exogenous time series variables and business hierarchy information.

Please cite this article as: C.S. Bojer and J.P. Meldgaard, Kaggle forecasting competitions: An overlooked learning opportunity. International Journal of Forecasting (2020), https://doi.org/10.1016/j.ijforecast.2020.07.007.

performance of methods and created a model selection framework based on the forecast horizon and five time series features comprising the seasonality, trend, cycle, randomness, and number of observations. The issue of "horses for courses" was explored further by Spiliotis et al. (2020) who examined whether the M competition data sets were diverse and representative of business forecasting, which is a prerequisite for using the data set for model selection. They used the feature-based instance space method proposed by Kang et al. (2017) to visualize the data sets in two-dimensional space and concluded that the M4 competition data sets were more diverse than those in previous competitions. However, they noted that the M4 competition data sets did not contain intermittent time series and they might have contained insufficient high-frequency series to guide model selection (Spiliotis et al., 2020).

The characteristics of the high-frequency time series included in the M4 competition have also been discussed. ForecastPro was the winner at the weekly frequency in the M4 competition and this team noted that many of the weekly time series differed from those 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 because this information can be used to identify the appropriate aggregation level for forecasting (Darin & Stellwagen, 2020). Similar points were made by Fry and Brundage (2020) who requested more high-frequency time series, hierarchy information, and access to potentially relevant exogenous variables. Based on their experience, they suggested that methods based on cross-learning, ML, and meta-models can provide significant improvements compared with statistical methods in these types of business forecasting tasks. In addition, they suggested that the lack of access to hierarchy information and exogenous variables were responsible for the poor performance of pure ML models. The lack of these data might also explain why the top two methods based on ML performed comparatively worse in terms of accuracy at the daily and weekly frequencies, although these frequencies typically have larger sample sizes than those at lower frequencies (Makridakis et al., 2020a).

To address these criticisms, the M Open Forecasting Center (2020) announced the M5 competition, which requires the forecasting of 42,840 daily time series of hierarchical sales data starting at the item level and aggregating to those of departments, product categories, and stores in three US states: California, Texas, and Wisconsin. In addition to time series data, the M5 competition includes external information such as price, promotions, day of the week, and special events, and the majority of the time series will exhibit intermittency. The competition is split into two parallel tracks using the same data set, where each has a cash prize of $50,000 USD. The first track 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 hosts the M5 competition and it has a large community of data scientists from a variety of backgrounds who 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 data set for a prediction task and typically offer a cash prize to the top performers. These competitions typically differ from academic competitions because they focus on solving a problem rather than learning why and when a particular method works. For example, 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 & Hyndman, 2011). Kaggle bases the final competition results on the private leaderboard performance, which is evaluated based on an unseen data set to prevent overfitting to the leaderboard and allow for ex-ante assessment.

3. Analysis of competitions

Initially, we examined the database of competitions from the online data science platform Kaggle, and we only retained the competitions focused on forecasting for further consideration, which resulted in nine forecasting competitions in the history of the platform. We decided to exclude the two earliest competitions comprising *Tourism Forecasting* and *GEFCOM 2012* because they were academically hosted competitions with published results and analyses (Athanasopoulos et al., 2011; Hong et al., 2014), which reduced the pool of competitions to the following seven competitions.

- Walmart Store Sales Forecasting[2]
- Rossmann Store Sales[3]
- Walmart Sales in Stormy Weather[4]
- Grupo Bimbo Inventory Demand[5]
- Wikipedia Web Traffic Time Series Forecasting[6]
- Corporación Favorita Grocery Sales Forecasting[7]
- Recruit Restaurant Visitor Forecasting[8]

All of these competitions featured high-frequency time series and access to hierarchy information, exogenous variables, or both. After a more thorough review of the data sets used for the competitions, we excluded the *Grupo Bimbo Inventory Demand* competition from further review because the data set contained a maximum of seven observations per time series at the weekly level

2 https://www.kaggle.com/c/walmart-recruiting-store-sales-forecasting.
3 https://www.kaggle.com/c/rossmann-store-sales.
4 https://www.kaggle.com/c/walmart-recruiting-sales-in-stormy-weather.
5 https://www.kaggle.com/c/grupo-bimbo-inventory-demand.
6 https://www.kaggle.com/c/web-traffic-time-series-forecasting.
7 https://www.kaggle.com/c/favorita-grocery-sales-forecasting.
8 https://www.kaggle.com/c/recruit-restaurant-visitor-forecasting.

Please cite this article as: C.S. Bojer and J.P. Meldgaard, Kaggle forecasting competitions: An overlooked learning opportunity. International Journal of Forecasting (2020), https://doi.org/10.1016/j.ijforecast.2020.07.007.

and many time series had only one observation, thereby making it unsuitable for time series forecasting.

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

The remaining two competitions both featured forecasts at a geographically disaggregate level, where daily forecasts of visits by webpage and restaurants were required, but the domains and number of time series differed. The *Recruit Restaurant* data set contained data regarding upcoming reservations for individual restaurants, which were expected to contain useful information for the forecasting task. The *Wikipedia* data set was a more traditional large-scale forecasting task, although it provided access to the geographical hierarchy through the page URL.

Thus, the competitions considered are varied in terms of their forecasting tasks, but they represented a more limited subset of the forecasting tasks conducted by companies compared with those present in the M4 competition. However, the competitions were more representative of business forecasting tasks with access to exogenous variables and the business hierarchy, which allowed us to examine the effects of these two factors.

3.1. Analysis of data sets

The goal of analyzing the identified Kaggle competition data sets was to position them relative to the M3 and M4 competitions in terms of simple time series characteristics such as entropy, seasonality, and trend. For this analysis, we utilized the methodology developed by Kang et al. (2017) to represent a single time series in two-dimensional space, thereby allowing the analysis of large-scale time series data sets.[9]

Data preprocessing. The Kaggle competition data sets were generally noisier than the M competition data sets, where most of the time series exhibited intermittence and others contained little historical data. Hence, the Kaggle competition data sets require some initial preprocessing to allow extrapolation of the time series instance space. We performed all of the preprocessing using the R packages **data.table** (Dowle & Srinivasan, 2019) and **base** (R Core Team, 2019). The preprocessing five steps are summarized as follows.

1. Set 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 sets.
4. Fill missing values in irregularly spaced time series[10] with zeroes.
5. Remove leading zeros.

Competition representativeness. Due to their ability to provide useful information about the M3 competition data, Kang et al. (2017) proposed the following 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 effects 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 whether the variance is approximately constant across the whole series.

To calculate the feature vectors, we used the R package **feasts** (O'Hara-Wild et al., 2019) and subsequently conducted principal component analysis to reduce the dimensionality by using the prcomp algorithm in the R package **stats** (R Core Team, 2019), thereby projecting them all into two-dimensional space to simplify their visualization with the R package **ggplot2** (Wickham, 2016). A similar method was also used by Spiliotis et al. (2020) in their assessment of the representativeness of the M4 competition. In their study, they used the *seasonal period (F4)* defined by the authors of the M4 competition, i.e., 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. Thus, seasonality cannot be estimated for weekly and daily series because the estimation algorithm requires at least two full seasonal periods.

To enable estimations 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 required 80 observations for the weekly series, and thus the seasonality could not be estimated for some of the series. However, this only affected around 20% of the time series for which we reverted to the original seasonal period of one and set the estimation of seasonality to zero for these observations. For the daily series, we

[9] Interested readers can refer to the complete analysis at https://github.com/cbojer/kaggle-project.

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

Please cite this article as: C.S. Bojer and J.P. Meldgaard, Kaggle forecasting competitions: An overlooked learning opportunity. International Journal of Forecasting (2020), https://doi.org/10.1016/j.ijforecast.2020.07.007.

Table 1
Selected Kaggle competitions and their associated forecasting tasks.

Competition	Time unit	Forecast unit	#Obser-vations	#Time series	Forecast horizon	Accuracy measure[a]	Exogenous 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[b]	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[b]	Store ID, Product ID, Department, Category, State	Interval

[a]See Appendix for the mathematical notations for the accuracy measures.
[b]SNAP refers to Supplemental Nutrition Assistance Program.

Please cite this article as: C.S. Bojer and J.P. Meldgaard, Kaggle forecasting competitions: An overlooked learning opportunity. International Journal of Forecasting (2020), https://doi.org/10.1016/j.ijforecast.2020.07.007.

selected a seasonal period of seven days because the *Restaurant Recruit* competition only contained 478 observations, thereby preventing estimation of the annual seasonality. Furthermore, only 65% of the daily time series in the M4 competition had more than two years of data.

Substituting the seasonal period of the weekly and daily time series made the *seasonal period (F4)* more influential in the dimensionality reduction process, where the time series with different seasonal periods became more dispersed because the distances were higher in terms of the *seasonal period (F4)* between the series with a period of one and a period of 52. Furthermore, we question the value of including the *seasonal period (F4)* as a numerical variable because it is essentially a categorical variable (hourly, daily, weekly, etc.) that is known in advance. Therefore, we consider that the impact of the seasonal period could be studied better by examining how the feature space of the frequencies differ rather than by including it as a feature. In the present study, the goal was to compare the M competitions with the Kaggle competitions and not to examine how the frequency affected the time series features. Therefore, we excluded the *seasonal period (F4)* from the dimensionality reduction process.

Fig. 1 shows the resulting time series instance spaces for the M3, M4, and Kaggle competitions, where the density of the data in each hexbin region is shown, with a low density in dark grey and a high density in blue. The figure highlights the differences between the M and Kaggle competition data sets. For the M competitions, the instance space was most densely populated on the right-hand side, thereby denoting strong levels for the *trend (F2)* and *ACF1 (F5)*. By contrast, all of the Kaggle competitions had density peaks further to the left, thereby indicating higher degrees of *entropy (F1)*. The M and Kaggle competitions data sets all included time series with varying *seasonality (F3)* and *lambda (F6)*. The discrepancies in terms of *trend (F2)*, *ACF1 (F5)*, and *entropy (F1)* can probably be explained by $\sim$ 95% of the time series in the M competitions being low frequency, i.e., either monthly, quarterly, or yearly, whereas all of the reviewed Kaggle competitions were high frequency, i.e., daily or weekly.

Fig. 1 also shows the similarities in terms of the positioning of time series with similar aggregation levels in the business hierarchy. The *Rossmann*, *Recruit Restaurant*, and *Walmart Store Sales* competitions were all at a high aggregation level and had density peaks within the same region of the time series instance space. The highly aggregated time series had lower degrees of *trend (F2)* and *ACF1 (F5)* but higher degrees *entropy (F1)* than the majority of the time series in the M4 competition. The *Corporacíon Favorita*, *Walmart Stormy Weather*, and *Wikipedia* competitions were all at low aggregation levels, but the similarity of their positions in the time series instance space was not clearly apparent. The *Corporacíon Favorita* and *Walmart Stormy Weather* competitions were similar in terms of their density peak areas, and they both exhibited relatively high degrees of *spectral entropy*, low degrees of *trend (F2)* and *ACF1 (F5)*, and varying degrees of *seasonality (F3)* and *lambda (F6)*. By contrast, the *Wikipedia* competition had higher levels for *trend (F2)* and *ACF1 (F5)* than the other competitions with low aggregation levels.

To some extent, we considered that the dissimilarity in *entropy (F1)* between the M competitions and the Kaggle competitions was caused by the selection criteria preventing intermittency in the M competitions because in contrast to the M competitions, all of the Kaggle competitions contained some intermittent[11] time series. In particular, more than 98% of the time series in the *Corporacíon Favorita*, *Walmart Stormy Weather*, *Recruit Restaurant*, and *Rossmann* competitions exhibited some degree of intermittency, and approximately 16% of the time series in the *Walmart Stores Sales* competition and 26% in the *Wikipedia* competition exhibited intermittency.

In summary, we found that the M competition data sets covered a significant part of the time series instance space, but most of these time series were characterized by a relatively high *strength of trend (F2)* and *first order autocorrelation (F5)* compared with those in the Kaggle competitions reviewed in this study. In addition, we found that the *spectral entropy (F1)* was the factor that differed between the M and Kaggle competition, particularly for the *Corporación Favorita* competition where the feature instance space even extended beyond that in the M4 competition. We consider that the dissimilarity in the *spectral entropy (F1)* was caused by the Kaggle competitions allowing intermittent series. Therefore, we argue that the inclusion of time series with higher degrees of *spectral entropy (F1)* and intermittence will improve the representativeness of future competition data sets.

3.2. Benchmarking Kaggle solutions

The fundamental basis of the present study is that the best performing solutions in the Kaggle competitions can provide valuable insights, which requires that the solutions should at least outperform the simple and proven time series forecasting methods. To verify whether this was the case, we benchmarked the solutions against two simple forecasting methods comprising the naïve and seasonal naïve methods. These methods are often used to identify whether a forecasting method or process adds value, such as in forecast value added analysis (Gilliland, 2011) or in terms of the mean absolute scaled error forecast accuracy measure (Hyndman & Koehler, 2006). We selected these two methods because they are simple and robust to missing data, which were present in all of the Kaggle competitions. Other frequently used benchmarking methods such as the theta and exponential smoothing models require time series without missing data. Therefore, these methods were not used because an imputation procedure would have been required to fill in the missing values before forecasting. However, methods such as theta or exponential smoothing can probably significantly outperform the benchmarks if carefully applied.

To construct the benchmark, we obtained 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 data set, and thus we had to use a fallback method where we

[11] Intermittence is used to describe the presence of zero observations in a time series.

Please cite this article as: C.S. Bojer and J.P. Meldgaard, Kaggle forecasting competitions: An overlooked learning opportunity. International Journal of Forecasting (2020), https://doi.org/10.1016/j.ijforecast.2020.07.007.

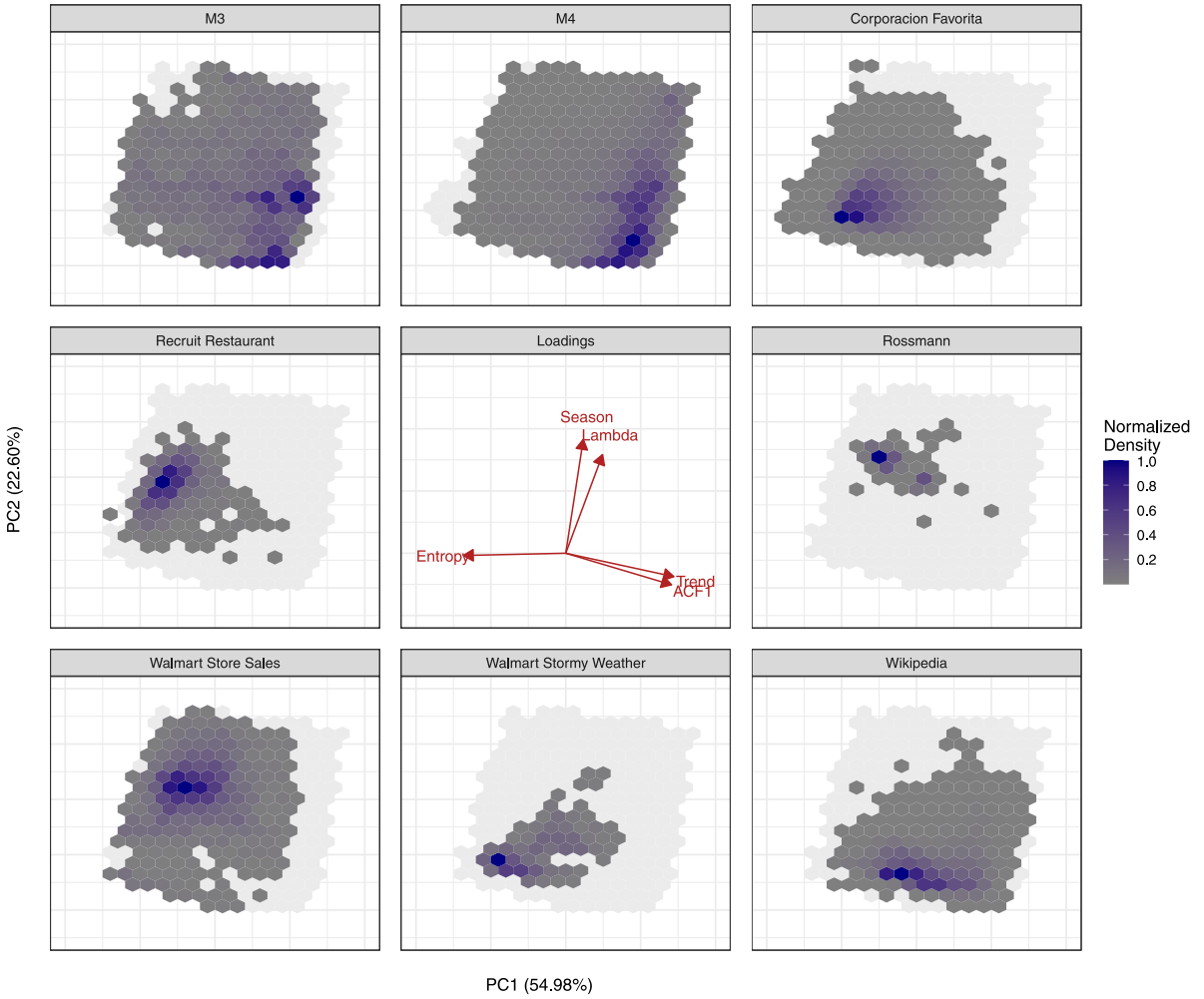

Fig. 1. Hexbin plots of the time series instance spaces for the M and Kaggle competitions. The color of each hexbin illustrates the density of the time series positioned in that particular field of the instance space, where blue denotes a high density and dark grey a low density. In addition, the instance space for the M4 competition is illustrated as a light gray background in all of the plots, except for the M4 plot, where the light gray background denotes the combined instance space for all competitions excluding the M4 competition. Furthermore, the *x*-axis and *y*-axis titles show the percentages of the total variance explained by the 1st and 2nd principal components, respectively. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

employed the mean at the next level of the forecast hierarchy to conduct simple cross-learning. In cases where data were still missing at the next level, we proceeded up the hierarchy until data were present.

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

$$\%Difference_{1st,nth} = \frac{Score_{nth} - Score_{1st}}{Score_{1st}} * 100,$$

where Score refers to the accuracy measure used in the competition. We used the same accuracy measure employed in each of the competitions because the selected accuracy measure reflected the business forecasting task and it was not possible to calculate other accuracy measures as the test set was not available. As a consequence, the differences were not directly comparable across competitions, e.g., a 20% difference in the root mean squared logarithmic error is not necessarily the same as a 20% difference in the weighted mean absolute error.

Fig. 2 shows the results obtained in the benchmarking procedures for the best of the two benchmarking methods and the top 25. Overall, the first place solutions provided improvements of greater than 25% compared with the simple benchmarks. The performance improvements were particularly striking for the *Rossmann* and *Corporación Favorita* competitions, where the benchmarks were more than 100% worse than the 1st place. Clearly, some competitions were much closer than others. In the *Corporación Favorita*, *Recruit Restaurant* and *Walmart Stormy Weather* competitions, the differences between the 1st and 25th places were relatively small, thereby indicating that any differences in performance could have been due to randomness. However, the differences were quite

Please cite this article as: C.S. Bojer and J.P. Meldgaard, Kaggle forecasting competitions: An overlooked learning opportunity. International Journal of Forecasting (2020), https://doi.org/10.1016/j.ijforecast.2020.07.007.

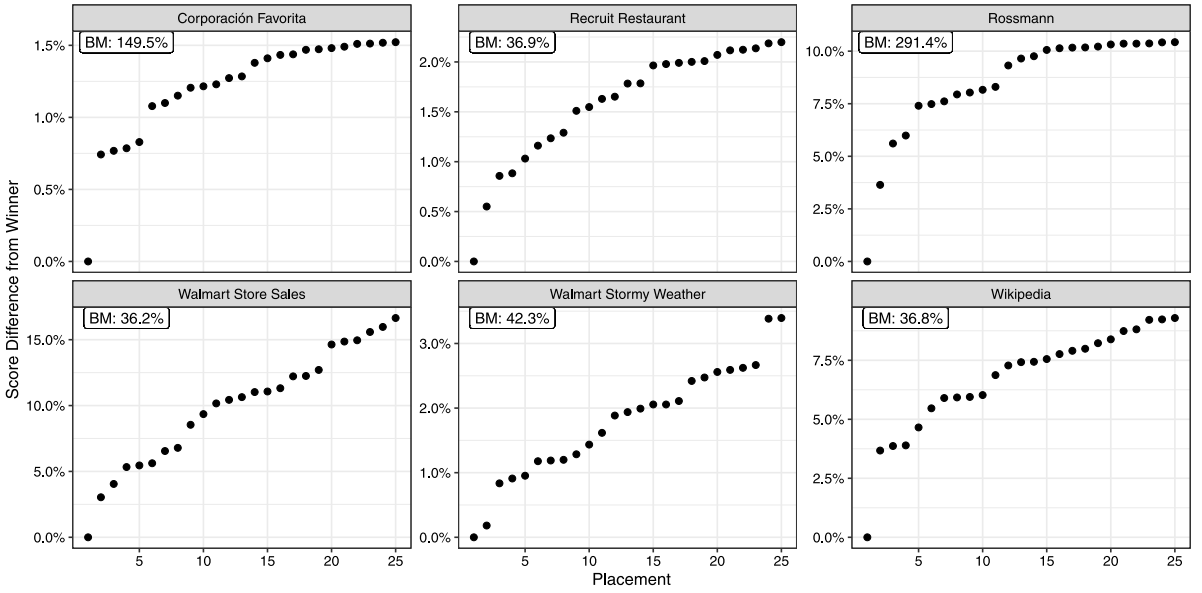

Fig. 2. Performance of the top 25 methods and the best benchmark in each competition. The performance was calculated as the percentage difference in terms of the score compared with the winner. The label in the top left corner of each subplot shows the performance of the best benchmark method.

significant in the other three competitions, thereby suggesting that the differences in performance at the top were meaningful. In these three competitions, the differences between the 1st and 2nd places were also more than 2.5%, which suggests that the winner's strategy was advantageous compared with that of the second place. In general, the benchmarks showed that the Kaggle solutions all added value compared with the simple time series benchmarks, thereby meriting further investigation.

4. Competition review

To conduct the review, we read through the Kaggle forum posts for each of the competitions. We gathered all information related to the solutions posted by the contestants, including both textual descriptions and code. Solutions in the top 25 were considered for review in each competition in order to focus on the top performers. In addition, the forums were examined to determine details about the application of simpler methods, such as historical averages or proven forecast methods, to investigate the improvements obtained compared with using simpler methods. Table 2 shows the reported solutions for the top 25 in each competition identified during the review process and the codifications for the methods used. Blank cells indicate contestants that did not describe their methods on the forums. Clearly, Table 2 shows that a significant proportion of the contestants did not report their methods, although they were usually reported by the top performers. This is a limitation of learning from the Kaggle competitions and we discuss the implications in Section 5.3. Table 2 also indicates another pattern where time series and statistical models mainly featured in the top 25 for the two Walmart competitions, whereas GBDT and neural networks mostly featured in the four later competitions. In the following sections, we present detailed findings based on our reviews of the six competitions.

4.1. Walmart Store Sales Forecasting (2014)

The Walmart Store Sales forecasting competition was the oldest of the competitions considered. This competition required contestants to provide weekly sales forecasts 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 the seasonality and holidays was found out to be crucial in this competition, where the 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 to then use these to denoise the individual time series. The winner accomplished this by conducting truncated singular value decomposition (SVD) for each category of time series (department), which was used to reconstruct the individual time series. Truncation had the effect of removing low-signal variations in the data to effectively filter out the noise. The denoised time series were then forecast using local forecasting methods such as "seasonal and trend decomposition using loess" (STL decomposition) combined with exponential smoothing and ARIMA. Finally, the forecasts from these methods were combined with an ensemble of simple forecasting models such as seasonal naïve, linear trend, and seasonality models, and historical averages. All of the ensembles improved the accuracy of the forecasts, but a single model from this ensemble comprising SVD

Please cite this article as: C.S. Bojer and J.P. Meldgaard, Kaggle forecasting competitions: An overlooked learning opportunity. International Journal of Forecasting (2020), https://doi.org/10.1016/j.ijforecast.2020.07.007.

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 was not available for the solution. 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, autoregressive integrated moving average (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), and **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)

followed by STL and exponential smoothing would have been sufficiently accurate to win the competition.

An important tweak used by all of the contestants in the top eight was to adjust the data to line up the holidays from year-to-year, which allowed them to be modeled as part of the seasonal pattern using time series models. ML models alone did not perform well in the competition and they were mostly used as parts of ensembles that also contained time series models. In particular, the second placed solution used a combination of ARIMA, unobserved components model, random forest, K-nearest neighbors, linear time series regression, and principal components regression for each department. Thus, the models were in the middle of the global vs. local dimension. Interestingly, this relatively complicated ensemble of models did not perform better than the simpler first place solution.

The exogenous variables available in the competition, including the temperature, fuel prices, CPI, unemployment, and information on markdowns, were not useful for producing accurate forecasts. Some of the top 10 contestants used these variables, but those in the top two and 4th place did not, thereby suggesting that they added little value. Other interesting contestants in the competition included the 3rd place due to its simplicity. This contestant lined up holidays and used a weighted average of the two closest weeks from the last year adjusted for the growth rate of the time series and warm days. This simple solution was only 4% worse than the best solution. Among the standard time series benchmarks, the simple naïve method was a strong benchmark but it was still beaten by more than 20% by all of the top 10 contestants.

4.2. Walmart Sales in Stormy Weather (2015)

The Walmart Sales in Stormy Weather competition had a slightly different format compared with the other competitions because the goal was to forecast the impact of extreme weather on sales. The aim of the competition was to provide daily unit sales forecasts by product and store for a total of 255 time series. The format differed from the other competitions because forecasts were not required for a future period. Instead, forecasts were required for a $\pm$ 3-day window around extreme weather

Please cite this article as: C.S. Bojer and J.P. Meldgaard, Kaggle forecasting competitions: An overlooked learning opportunity. International Journal of Forecasting (2020), https://doi.org/10.1016/j.ijforecast.2020.07.007.

event occurrences, which had been removed from the available data. Thus, this was not a forecasting task in the purest sense because 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 employed in retail forecasting software, which involves first estimating the baseline sales and then modeling the deviations from the baseline using linear regression with exogenous variables. The solution utilized projection pursuit regression with only time as an input to estimate the baseline sales per time series as well as considering the 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). Thus, the main difference from the typical approach for retail forecasting software was the use of a more complex smoother than the frequently employed moving average, as well as a global rather than local regression model. The winner constructed several features[12] from the exogenous variables, including modeling the weekend/weekdays, holidays, and their interactions, as well as time information (year, month, day, and trend) and modeling Black Friday (including lag and lead effects). As expected, weather data were used in the solution in the form of indicator variables to model the threshold effects for precipitation and departure from normal temperatures. However, in the description of their solution, the winner mentioned that using weather information did not improve the forecasting performance greatly, which was supported by the comments of other highly placed contestants. This finding is somewhat surprising because the aim of the competition was to predict the effects of extreme weather on sales and the actual weather was available instead of a forecast.

The top contestants used several other approaches, such as local Gaussian process regression by the 5th place solution with mainly date features. Ensembles of various ML models with models such as GBDT using the XGBoost algorithm (Chen & Guestrin, 2016), random forest, support vector machines, and linear regressions were employed successfully by the 3rd, 6th, and 11th place contestants. However, it should be noted that these complex ensembles of models, which generally performed well in later Kaggle competitions, did not perform better than the much simpler approach of the winner. This competition also included the first use of XGBoost, which finished in third place and its performance was not dominant. In addition, conventional time series models were not used frequently in the reported solutions, except the 6th place solution that employed time series models such as ARIMA as part of an ensemble. However, ARIMA did not achieve impressive performance on its own, where its performance was 17% worse than the winning solution on the public leaderboard.

[12] In this review, we use the term features to refer to model inputs, which comprised potentially processed external information and/or time series information.

4.3. Rossmann Store Sales (2015)

The Rossmann Store Sales competition was characterized by the increased use of ensembles of global ML models, particularly XGBoost. This was also the first competition where a neural network placed in the top three. The competition required that contestants 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 the other contestants mainly by adapting the XGBoost model to perform well based on time series data. 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 address the fact that GBDT cannot extrapolate trends. The main innovations among the features comprised calculating statistics and their rolling versions at different levels of the hierarchy and for different days of the week and promotion periods, where examples include the average sales by product, moving averages of sales by product, and average sales by product and promotion status. In addition, event counters proved useful, which comprised the number of days until, during, and after an event, such as holidays or promotions. The solution also included weather information in the form of precipitation and the maximum temperature together with seasonality indicators, including the month, year, day of month, week of year, and day of year, to facilitate accurate estimations of multiple seasonal effects. The key to good performance by many ML models was the appropriate selection of features and hyperparameters to maximize the accuracy but without overfitting the training data set. The strategy used by many contestants involved utilizing a hold-out data set with the same length as the forecast horizon to evaluate the quality of the model and to determine the hyperparameters and select features. The use of ensembling multiple XGBoost models improved the performance by around 5% compared with the best single model. Variation was introduced into the ensemble by training the model based on different data subsets, training models using both direct and iterated predictions, and including different subsets of features in the models.

Most of the top performers used ensembles of global XGBoost models to produce forecasts, but some included local XGBoost models as part of their ensemble. The features used were generally similar to those employed by the winner, where they included event counters and statistics calculated at various levels of the hierarchy. Thus, the exogenous variables related to events, i.e., holidays and promotions, were essential for obtaining high performance in this competition, and they probably explain the significant performance improvements compared with the seasonal naïve benchmark. The two best solutions were distinguished by the use of rolling statistics in the form of moving averages or medians as features, thereby adapting and utilizing well-known time series forecasting methods.

Please cite this article as: C.S. Bojer and J.P. Meldgaard, Kaggle forecasting competitions: An overlooked learning opportunity. International Journal of Forecasting (2020), https://doi.org/10.1016/j.ijforecast.2020.07.007.

The 3rd place solution successfully applied neural networks for the first time in the forecasting competitions on Kaggle. The neural network employed 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 was learned and utilized by the network for prediction. The solution did not include autoregressive inputs, although this is usually the case for neural networks in forecasting. We refer the reader to the report posted by the contestants for further details (Guo & Berkhahn, 2016).

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

4.4. Wikipedia Web Traffic Forecasting (2017)

The Wikipedia Web Traffic Forecasting competition took scale to another level, as it required forecasts for more than 145,000 time series. This competition also showcased the power of deep learning for forecasting, which won the competition and occupied six places in the top eight. The competition required that contestants 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 used an elegant and accurate deep learning approach without much feature engineering,[13] which differs from the solutions using GBDT. The solution comprised an ensemble of global recurrent neural networks with identical structures. Multiple ensembling approaches were used to reduce the variance in the predictions because neural network predictions can be volatile with noisy data. Three models were trained based on different random seeds to counteract the randomness of the weight initializations in the network. Two approaches were used to prevent sensitivity to the exact number of training iterations conducted. First, model checkpoints were saved during the training procedure and the averages of the checkpoint predictions were used. Second, moving averages of the neural network weights were used instead of the final weights, which is also known as stochastic weight averaging (Izmailov et al., 2018). The features used in the neural networks comprised historical page views and categorical variables, such as the agent, country, site, and day of the week. One weakness of recurrent neural networks is that they have difficulty modeling long-term dependencies, such as yearly seasonality. The winner addressed this problem by including the page views from a quarter, half-year, and year ago as inputs in the model. They also included the autocorrelation function value at lag 365 and lag 90 to improve the modeling of the yearly seasonality. Series were independently scaled to facilitate cross-learning the seasonality and time dynamics, and 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 employed together with an automated hyperparameter tuning algorithm by using the Bayesian optimization algorithm SMAC3 (Lindauer et al., 2017) to determine the hyperparameters of the neural networks. Interestingly, the winner reported that the final performance was relatively insensitive to the hyperparameters and the algorithm obtained several models with similar performance.

The other top performers used different neural network architectures, including recurrent neural networks, convolutional neural networks, and feedforward neural networks, thereby demonstrating that several different architectures could obtain similar performance. In addition, feature engineering was used by the top contestants to varying degrees. The 4th and 6th place solutions used limited feature engineering, whereas the 2nd place solution conducted extensive feature engineering, including employing the predictions from various ensemble models as inputs for another model (referred to as stacking in ML). Thus, many different architectures produced good solutions and complex feature engineering was not a requirement for high performance neural network forecasts with this data set. Neural networks dominated the competition but another much simpler solution in 8th place involved a segmented approach, which included Kalman filters to predict high signal series and a robust approach based on the median of the moving medians over different windows to predict low signal series. This solution was the only approach in the top 10 to employ traditional time series models and although it performed well, the solution was still about 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 upon the solutions obtained in previous competitions because both the gradient boosting approaches used in the *Rossmann* competition and the neural network approaches applied in the *Wikipedia* competition were utilized greatly by the top

[13] Feature engineering refers to the construction of features from exogenous variables or the time series itself.

performing contestants. The competition required that contestants 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 comprising both gradient boosting models and neural network models. One change from the 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 innovative feature of the solution involved training one model per forecast horizon rather than one model for all forecast horizons in order to allow the models to learn the useful information for each horizon. This approach yielded good results but there was a trade-off due to the requirement for 16 models rather than one model. This approach was used for a LightGBM model and a feedforward neural network in an ensemble with two other models, where these models comprised another LightGBM model trained for all horizons and the convolutional neural networks 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 employed included measures of centrality and spread, as well as an exponential moving average.

Interestingly, the winner only used very recent data in their models and they preferred to drop older observations based on the performance of the validation data set. 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 placed solutions also favored this approach, such as those in the 5th and 6th place. This approach may have worked despite ignoring the yearly seasonality due to the trend present in the data and the short forecast horizon of only 16 days.

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

A hold-out strategy was used in most of the previous competitions to prevent overfitting but several contestants experimented with other validation approaches. For example, the 4th place solution held out a certain percentage of the time series, thereby relying only on cross-learning for performance estimation. Another interesting validation approach involved the use of a combination of grouped K-fold cross-validation to estimate parameters and time series cross-validation to estimate the model performance. In the grouped K-fold cross-validation, each time series was restricted to one fold to avoid information leakage across folds, thereby also relying only on cross-learning. The time-series cross-validation used two consecutive hold-out data sets of 16 days to estimate the model performance. Despite these interesting validation approaches working successfully for forecasting, the hold-out approach seems to continue to suffice because it was employed by the top three solutions.

4.6. Recruit Restaurant Visitor Forecasting (2018)

The Recruit Restaurant Visitor Forecasting competition confirmed the success of previously used methods such as GBDT using rolling statistics, and neural networks to some degree, in a different domain. The competition required that contestants 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 of the competition was a team of four contestants who used an ensemble comprising 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 for the restaurant reservations, which were the main differences from earlier competitions in different domains. Another challenge in the competition was that the test set included the "Golden Week" holiday period with significantly different behavior, and the contestants only had access to one earlier holiday period in the training data set. 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 before and after as Fridays and Mondays, respectively. In general, this tweak significantly improved performance as evaluated after the competition by multiple contestants in the top places. The use of this tweak was not necessary to win the competition, as demonstrated by the 1st place solution. However, it highlights the value of using domain knowledge and manual adjustments to the data to achieve the best possible performance in a similar manner to the findings obtained in the Walmart Store Sales competition.

The 1st and 5th place solutions used neural networks, but they were generally not as successful compared with earlier competitions and they were mainly employed 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 contestants in the 21st, 23rd, and 25th places utilized these methods and their accuracy was around 2% worse than that of the contestant in the 1st place. The results may have been affected by the size of the data set, which was smaller than the *Wikipedia* and *Corporación Favorita* data sets by a factor over 100. Interestingly, a Kalman filter managed to place competitively in 33rd place in a similar manner to the *Wikipedia* competition and the difference in performance was only 2.4%

Please cite this article as: C.S. Bojer and J.P. Meldgaard, Kaggle forecasting competitions: An overlooked learning opportunity. International Journal of Forecasting (2020), https://doi.org/10.1016/j.ijforecast.2020.07.007.

compared with that in 1st place, thereby demonstrating that more traditional time series models were still viable when exogenous variables were available.

Similar to the earlier competitions, most contestants used a hold-out data set for model performance validations, although it was surprising that both the contestants in 7th and 8th places managed to perform well using a standard K-fold validation approach while ignoring the time series nature of the data. Inspired by the innovation in the *Corporación Favorita* competition, some contestants trained horizon-specific models, where the solutions ranked in 1st place and 5th place required a total of 42 models. A compromise was made by the contestant in 11th place who trained one model per week for six models in total but they still modeled some of the potential horizon-specific effects. However, not all of the top ranked contestants used horizon-specific models, thereby suggesting that the performance improvements obtained with this approach might not be substantial compared with the growth in the number of models.

5. Discussion

In this section, we consider the insights obtained 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 insights obtained;
- Providing nine ex-ante forecasts for the outcomes of the M5 competition;
- Discussing the limitations of the insights obtained 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 this was the case across different domains and forecasting tasks. Cross-learning was also used by all of the competition winners, although sometimes in combination with local models, thereby highlighting the benefits of cross-learning for time series and motivating further research in this area.

External information. The differences in performance between global and local models with the benchmarks suggest that access to the business hierarchy provides even greater cross-learning benefits than those found in the M4 competition. Access to exogenous variables other than the hierarchy provided substantial benefits in some competitions but very small or no benefits in others. Available information known in advance, such as promotions, holidays, events, and reservations, was highly useful in most of the competitions. However, variables that need to be forecast, such as weather and macroeconomic variables, appeared to provide no significant benefits despite the availability of actual values rather than forecasts in the competitions reviewed. A similar finding was obtained in the non-public part of the *Tourism Forecasting* competition (Athanasopoulos et al., 2011), which suggests that this applies in multiple domains.

Statistics vs. ML. In our review of the six competitions, no single method dominated all of the competitions. The two earliest competitions, *Walmart Store Sales* and *Walmart Stormy Weather*, were won by innovative applications 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. In addition, surprisingly similar structures were employed 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? A possible answer is that these methods were not sufficiently mature or well 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 was unimpressive in the NN3 competition (Crone et al., 2011). The key to success with neural networks appears to be the use of cross-learning and adopting various innovative architectures, such as the long short-term memory model (Hochreiter & Schmidhuber, 1997) and embedding layers (Guo & Berkhahn, 2016). The first successful GBDT algorithm comprising XGBoost was not released until after the first Walmart competition. XGBoost was available and used in the *Walmart Stormy Weather* competition, but the method was still new and adaptations to the time series domain were not developed. Therefore, it is impossible to determine whether the competitions would still be won by time series and statistics methods if they were held today.

A better question might be why did time series and statistical methods not perform competitively in the latest four competitions? We consider that the characteristics of the data sets in the last four competitions were better suited to both GBDT and neural networks. The four latest data sets were all characterized by intermittency and they contained external information relevant to the forecasting task in the form of hierarchy information and predictive exogenous variables, such as holidays, events, promotions, and reservations. By contrast, the *Walmart Store Sales* competition was continuous with access to hierarchy information, and the exogenous variables contained little useful information, probably due to the high aggregation level, thereby providing ideal conditions for global time series methods. The *Walmart Stormy Weather* competition had the smallest data sets and little business hierarchy information, which limited the opportunity for cross-learning. The only exogenous variables provided were related to weather, which were found to be not very useful. Furthermore, the availability of data from before and after the required forecasting periods and the short forecast horizon made this competition ideal for statistical smoothing methods, such as the projection pursuit regression approach employed by the winner. Thus, we found that for disaggregated data sets characterized by

Please cite this article as: C.S. Bojer and J.P. Meldgaard, Kaggle forecasting competitions: An overlooked learning opportunity. International Journal of Forecasting (2020), https://doi.org/10.1016/j.ijforecast.2020.07.007.

intermittency and containing relevant external information, ML methods performed better than both time series and statistical methods, which agrees with the practical experience of forecasters at both Google (Fry & Brundage, 2020) and Amazon (Salinas et al., 2020).

A frequent concern regarding more complex methods is their practical applicability and whether the potential accuracy gains might justify the added complexity and computational requirements (Gilliland, 2020). The ML methods used in the four most recent Kaggle competitions all required the training of multiple complex models, and thus they were more expensive than the popular time series benchmarks in terms of their cost and time. In this review, we found that the best solutions generally obtained considerable improvements compared with the simple benchmarks, where the seasonal naïve method obtained results between 35% and 290% worse than the winners. Thus, the use of more complex methods that effectively employ the business hierarchy and exogenous variables should be seriously considered for daily and weekly business forecasting tasks, and further research should investigate this trade-off dimension in more detail, e.g., by benchmarking the ML methods vs. exponential smoothing and other proven statistical methods.

The GEFCOM 2014 and 2017 competitions were from an entirely different domain compared with the competition reviewed in this study, but they also featured high-frequency time series, access to exogenous variables, and competitors that used both statistics and ML methods. In the GEFCOM competitions, statistics or ML methods did not emerge as clear winners across the competitions in a similar manner to our findings, which is 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 suggest that this was due to differences in the characteristic of the data sets employed in the five competitions. We note that despite these differences, methods from both statistics and ML were utilized by at least one top ranked competitor in all of the GEFCOM competitions. The load forecasting competitions won by statistical methods were characterized by relatively strong seasonality and a well-understood relationship with one key variable, i.e., temperature. ML methods performed better in the wind and solar power forecasting challenges characterized by intermittency, a lower signal-to-noise ratio, and weaker relationships with the exogenous variables. Further studies should investigate the relationships between the performance of ML and statistical methods, and the data set characteristics in more detail. We suggest considering intermittency and the information content in terms of both exogenous variables and the business hierarchy as important factors. Thus, a key question is how to operationalize the concept of information content in a manner that generalizes across time series data sets.

Gradient boosted decision trees vs. neural networks. Regarding the differences in performance between GBDT and neural networks, we found that neural networks outperformed GBDT in the Wikipedia competition, which used a very large data set that contained no useful exogenous variables. In the other three of the latest competitions, both methods were ranked at the top. Therefore, we suggest that the strength of GBDT is its ability to model external information. Neural networks have been employed widely in recent forecasting research and they were used by the winner of the M4 competition (Smyl, 2020). However, we are not aware of any previous research into using GBDT in combination with the strategies from the Kaggle competitions. The second place solution in the M4 competition used GBDT but it was employed as a meta-learner to combine traditional time series forecasting methods (Montero-Manso et al., 2020).

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

Validation strategies. In all six of the competitions, we found that a hold-out data set with a length equal to the forecast horizon was employed successfully to validate the model's performance and prevent overfitting. It is somewhat surprising that we did not find substantial overfitting to the validation set when it was used for multiple evaluations of ML models to select features and hyperparameters. A potential explanation is the public leaderboard feedback provided by the Kaggle platform because a performance drop on the leaderboard would indicate that contestants are probably overfitting the validation set. Therefore, the approach adopted by many contestants generally corresponds to splitting the data in the following four ways:

1. A training set for estimating the models;
2. A validation set for evaluating the model's performance and performing model diagnostics;
3. A second small validation set where only the summary performance measure is available to prevent overfitting;
4. The final test set used for evaluating the 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, where it can facilitate learning and help avoid overfitting if used appropriately. However, contestants that relied only on the public leaderboard for validation often found that they were ranked lower on the private leaderboard due to overfitting.

Please cite this article as: C.S. Bojer and J.P. Meldgaard, Kaggle forecasting competitions: An overlooked learning opportunity. International Journal of Forecasting (2020), https://doi.org/10.1016/j.ijforecast.2020.07.007.

The fact that Kaggle does not make the public leaderboard data available for retraining models forces contestants to hold out the most recent data for validation, and thus the model has to forecast further compared with alternative validation strategies. The M5 competition has addressed this issue by providing contestants with access to the public leaderboard data before forecasts are required for the final test set. Further studies should evaluate how the inclusion of a leaderboard that is revealed later compares with 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 data set from the retail domain, which was generously supplied by Walmart. The competition will require forecasts for more than 40,000 daily time series at the store and product level, and contestants will be provided with information on the prices, promotions, events, and product hierarchy (M Open Forecasting Center, 2020). Thus, the forecasting task is very similar to that in the *Corporación Favorita* competition. The main difference compared with the competition reviewed in the present study is the need to evaluate the prediction uncertainty as well as the prediction accuracy. Based on the insights obtained from our review, we provide the following hypotheses formulated prior to the start of the M5 competition.

- The instance space representation of the time series in the M5 competition will resemble that in the *Corporación Favorita* competition, and thus the entropy and trend are higher, and the first-order autocorrelation is lower compared with the 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 difference in performance between local models and models that use cross-learning compared with the M4 competition.
- GBDT using feature engineering based on methods such as rolling statistics and neural networks will perform well in the competition, and outperform the existing time series benchmarks in terms of both the 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, as addressed in many recent studies (e.g., see Duan et al. (2019) for GBDT and Salinas et al. (2020) for neural networks).
- Ensembles of methods will continue to occupy the top ranks, which is consistent with the findings obtained in all of the Kaggle and M competitions. We expect that these ensembles will contain both neural networks and GBDT, potentially in combination with other methods.
- Hold-out data sets or time series cross-validation will be used by the top ranked contestants to avoid overfitting.
- Using exogenous variables that are known in advance, such as prices, promotions, holidays, and other events, will help to improve the forecast accuracy, as shown in previous retail research (Fildes et al., 2019) and our review of the Kaggle competitions.
- Contestants will develop innovative strategies to address the challenge of hierarchical forecasting, and we expect that new neural network architectures and GBDT strategies will be employed to utilize this information in an optimal manner.

5.3. Limitations

The focus on providing solutions to real-life forecasting tasks in the competitions reviewed in this study has a disadvantage because there are limitations regarding what researchers can infer from the competitions. Lack of access to the test set after the competition has ended means that it is impossible to test for significant differences in the performance of various solutions or to evaluate performance using alternative error measures. Furthermore, it is not possible to analyze the performance of different solutions based on various subsets of the data set to improve our understanding of the strengths and weaknesses of various methods. Future Kaggle competitions should address this issue by making the test set available after the competition has ended in the same manner as the M4 and M5 competitions.

A major weakness in terms of the practical applicability of the Kaggle competitions reviewed in this study is that they failed to address the prediction uncertainty. Forecasts are always wrong, and thus an estimate of the uncertainty associated with the forecast is crucial for decision making based on forecasts, e.g., in hedging, capacity planning, and inventory management. The competitions were based on real-life forecasting tasks, so it is surprising that the competition case companies did not demand prediction intervals as part of the forecasting task, although it is possible that the companies do not 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 future Kaggle competitions.

Kaggle encourages the sharing of solutions but contestants are not required to share their solution or code publicly, which makes it harder to learn from competitions and it does not allow the reproduction of results. Ideally, Kaggle should require that contestants submit their code, thereby facilitating the reproducibility of results as well as full mapping and analysis of the solutions. A less strict alternative might involve asking contestants to complete a small survey containing questions about the methods they employed when submitting their final predictions. This option will not address the issue of reproducibility but it would facilitate learning from competitions by enabling the mapping of the solutions,

Please cite this article as: C.S. Bojer and J.P. Meldgaard, Kaggle forecasting competitions: An overlooked learning opportunity. International Journal of Forecasting (2020), https://doi.org/10.1016/j.ijforecast.2020.07.007.

although it will necessarily be less detailed than what could be possible if code was available.

The lack of publicly-shared solutions also has implications for the validity of our review. It is possible that the use of methods such as linear regression or local time series models in the non-reported solutions in the top 25 places would change our results. However, we consider it highly unlikely that local time series methods would have performed competitively in the four latest competitions due to the intermittency and influence of exogenous variables in the data sets. Our benchmarking results also support this conclusion. We also consider that the presence of a systematic reporting bias caused by differences in the willingness to share different solution methods was unlikely. Despite these weaknesses, we still believe that much can be learned by focusing on the patterns that worked across the competitions and determining the relationships between our findings and the characteristics of the data sets. Further studies should test our hypotheses with a variety of data sets, and the upcoming M5 competition will certainly serve as a suitable initial testbed.

6. Conclusions

Based on our analysis and review of six recent Kaggle forecasting competitions, we consider that the forecasting community has much to learn from the Kaggle community in terms of forecasting daily and weekly business time series. Our analysis showed that the M4 competition data set contained similar time series to those used in the Kaggle competitions, although time series with these characteristics were underrepresented in the M4 competition data set. Furthermore, the Kaggle data sets differed from the M4 competition because they provided access to external information, e.g., exogenous variables or business hierarchy, which led to significant improvements in the forecast accuracy.

Similar to the findings obtained from the M4 competition, we showed that global ensemble models outperformed local single models. In contrast to the M4 and two earlier Kaggle competitions, conventional time series and statistical methods were significantly outperformed by ML methods in the four latest Kaggle competitions, which may be attributed to the utilization of external information by the ML methods to cross-learn and model the effects of exogenous factors. In addition, we found that the top ranked solutions in the Kaggle competitions and the top two solutions in the M4 competition were similar, where they relied on either GBDT or neural networks. However, several adaptations must be made to the ML methods and their validation strategies to obtain performance benefits from the ML methods.

We strongly encourage the forecasting community to learn from the reviewed ML strategies for time series forecasting and to participate in their further development. The M5 competition provides an ideal opportunity because the forecasting task and data set are very similar to those used in some of the Kaggle competitions reviewed in this study. Therefore, we consider that our hypotheses based on the Kaggle competitions discussed in this study will foretell the results of the M5 competition.

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.

Appendix. Competition accuracy measures

For the accuracy measures, h denotes the forecast horizon, y_i are the actual values, $\hat{y}_i$ are the forecast values, and w_i is the weight assigned to observation i. The weight, w_i, is used to penalize forecast errors for particular observations or time series, i.e., 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)

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

References

Athanasopoulos, G., & Hyndman, R. J. (2011). The value of feedback in forecasting competitions. *International Journal of Forecasting*, *27*(3), 845–849. http://dx.doi.org/10.1016/j.ijforecast.2011.03.002, URL http://www.sciencedirect.com/science/article/pii/S0169207011000495.

Please cite this article as: C.S. Bojer and J.P. Meldgaard, Kaggle forecasting competitions: An overlooked learning opportunity. International Journal of Forecasting (2020), https://doi.org/10.1016/j.ijforecast.2020.07.007.

Athanasopoulos, G., Hyndman, R. J., Song, H., & Wu, D. C. (2011). The tourism forecasting competition. *International Journal of Forecasting*, *27*(3), 822–844. http://dx.doi.org/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* (pp. 785–794). Association for Computing Machinery, http://dx.doi.org/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*(3), 635–660. http://dx.doi.org/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*(1), 135–141. http://dx.doi.org/10.1016/j.ijforecast.2019.03.018, URL http://www.sciencedirect.com/science/article/pii/S0169207019301177.

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. http://arxiv.org/abs/1910.03225.

Fildes, R. (2020). Learning from forecasting competitions. *International Journal of Forecasting*, *36*(1), 186–188. http://dx.doi.org/10.1016/j.ijforecast.2019.04.012.

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

Fildes, R., & Ord, K. (2007). Forecasting competitions: Their role in improving forecasting practice and research. In M. P. Clements, & D. F. Hendry (Eds.), *A companion to economic forecasting* (pp. 322–353). John Wiley & Sons, Ltd, 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*(1), 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*(1), 161–166. http://dx.doi.org/10.1016/j.ijforecast.2019.04.016, URL http://www.sciencedirect.com/science/article/pii/S0169207019301165.

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. *The Journal of the Operational Research Society*, *20*(4), 451–468. http://dx.doi.org/10.1017/cbo9780511753961.021.

Guo, C., & Berkhahn, F. (2016). Entity embeddings of categorical variables. arXiv:1604.06737 Cs.

Hibon, M., & Makridakis, S. (2000). The M3-Competition: results, conclusions and implications. *International Journal of Forecasting*, *16*(4), 451–476.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, *9*(8), 1735–1780. http://dx.doi.org/10.1162/neco.1997.9.8.1735.

Hong, T., Pinson, P., & Fan, S. (2014). Global energy forecasting competition 2012. *International Journal of Forecasting*, *30*(2), 357–363. http://dx.doi.org/10.1016/j.ijforecast.2013.07.001.

Hyndman, R. J. (2020). A brief history of forecasting competitions. *International Journal of Forecasting*, *36*(1), 7–14. http://dx.doi.org/10.1016/j.ijforecast.2019.03.015, URL http://www.sciencedirect.com/science/article/pii/S016920701930086X.

Hyndman, R. J., & Koehler, A. B. (2006). Another look at measures of forecast accuracy. *International Journal of Forecasting*, *22*(4), 679–688. http://dx.doi.org/10.1016/j.ijforecast.2006.03.001, URL http://www.sciencedirect.com/science/article/pii/S0169207006000239.

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*(2), 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* (pp. 3149–3157). Curran Associates Inc.

Lindauer, M., Eggensperger, K., Feurer, M., Falkner, S., Biedenkapp, A., & Hutter, F. (2017). *SMAC v3: Algorithm configuration in Python*. GitHub, 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. *International Journal of Forecasting*, *36*(1), 29–36. http://dx.doi.org/10.1016/j.ijforecast.2019.02.012.

Makridakis, S., Spiliotis, E., & Assimakopoulos, V. (2020b). Responses to discussions and commentaries. *International Journal of Forecasting*, *36*(1), 217–223. http://dx.doi.org/10.1016/j.ijforecast.2019.05.002, URL http://www.sciencedirect.com/science/article/pii/S0169207019300871.

Makridakis, S., Spiliotis, E., & Assimakopoulos, V. (2020c). The M4 Competition: 100,000 time series and 61 forecasting methods. *International Journal of Forecasting*, *36*(1), 54–74. http://dx.doi.org/10.1016/j.ijforecast.2019.04.014, URL http://www.sciencedirect.com/science/article/pii/S0169207019301128.

Montero-Manso, P., Athanasopoulos, G., Hyndman, R. J., & Talagala, T. S. (2020). FFORMA: Feature-based forecast model averaging. *International Journal of Forecasting*, *36*(1), 86–92. http://dx.doi.org/10.1016/j.ijforecast.2019.02.011, URL http://www.sciencedirect.com/science/article/pii/S0169207019300895.

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*(1), 3–6. http://dx.doi.org/10.1016/j.ijforecast.2019.05.005, URL https://www.sciencedirect.com/science/article/pii/S0169207019301116.

Petropoulos, F., Makridakis, S., Assimakopoulos, V., & Nikolopoulos, K. (2014). 'Horses for Courses' in demand forecasting. *European Journal of Operational Research*, *237*(1), 152–163. http://dx.doi.org/10.1016/j.ejor.2014.02.036, URL https://www.sciencedirect.com/science/article/pii/S0377221714001714.

R Core Team (2019). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation 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*(3), 1181–1191. http://dx.doi.org/10.1016/j.ijforecast.2019.07.001, URL http://www.sciencedirect.com/science/article/pii/S0169207019301888.

Smyl, S. (2020). A hybrid method of exponential smoothing and recurrent neural networks for time series forecasting. *International Journal of Forecasting*, *36*(1), 75–85. http://dx.doi.org/10.1016/j.ijforecast.2019.03.017, URL http://www.sciencedirect.com/science/article/pii/S0169207019301153.

Spiliotis, E., Kouloumos, A., Assimakopoulos, V., & Makridakis, S. (2020). Are forecasting competitions data representative of the reality?. *International Journal of Forecasting*, *36*(1), 37–53.

Vowpal Wabbit (2007). *Vowpal Wabbit*. GitHub, URL https://github.com/VowpalWabbit/vowpal_wabbit.

Wickham, H. (2016). *ggplot2: Elegant graphics for data analysis*. New York: Springer-Verlag, URL https://ggplot2.tidyverse.org.

Please cite this article as: C.S. Bojer and J.P. Meldgaard, Kaggle forecasting competitions: An overlooked learning opportunity. International Journal of Forecasting (2020), https://doi.org/10.1016/j.ijforecast.2020.07.007.