The value of feedback in forecasting competitions

George Athanasopoulos and Rob J Hyndman

10 March 2011

Abstract

In this paper we challenge the traditional design used for forecasting competitions. We implement an online competition with a public leaderboard that provides instant feedback to competitors who are allowed to revise and resubmit forecasts. The results show that feedback significantly improves forecasting accuracy.

1 Introduction

Over the past four decades, forecasting competitions have proven to be a valuable tool for evaluating the accuracy of forecasting methods. Some of the earliest works in the area of economics and business date back to Reid (1969) and Newbold & Granger (1974). These were followed by the series of "M-competitions" by Makridakis, Hibon and co-authors, with the Tourism Forecasting Competition (TFC) of Athanasopoulos et al. (2011) being the latest contribution in this line of research. These competitions have played a significant role in the advancement of forecasting practices. However, the conduct of the competitions has not seen a similar development.

Typically, a competition is carried out in three stages. First, data are collected and forecasting tasks are set; second, forecasts are generated; and finally, forecast evaluation results are presented and analysed. In this setting the forecaster is never presented with the opportunity to evaluate and revise the forecasting methods implemented. Revising, updating and adjusting forecasting methods is common in practice, but it has not been part of these forecasting competitions. In this paper we describe an online competition using the data from the TFC. The competition utilized a public "leaderboard" allowing participants to evaluate their forecasts, learn from the results, and to revise and resubmit new forecasts.

2 The online competition

The TFC mainly evaluated the forecasting performance of "off the shelf" popular methods commonly used in practice. After our paper was completed, we decided to extend the competition to allow other contributions via Kaggle (www.kaggle.com. Kaggle was established in 2010 and is the only web-based platform for data prediction competitions. It provides a means for its clients to expose prediction challenges they face to a wide range of forecasters. Kaggle forecasting participants have very diverse backgrounds ranging from computer science, statistics and mathematics to engineering and physics and come from over 100 countries and 200 universities. We thought the online extension of the competition might allow more tailored methods and lead to new insights.

We implemented the online competition in two stages: stage one involved forecasting only the yearly data (www.kaggle.com/tourism1), and stage two involved forecasting the monthly and the quarterly data (www.kaggle.com/tourism2). This was done to prevent participants from attempting to forecast the yearly data by aggregating forecasts generated for the higher frequency data as this was a successful strategy in the TFC. The winning entry would be the one that scored the lowest MASE (calculated as in the TFC), averaged across monthly, quarterly and yearly data series, for forecast horizons $h = 1$ to 24, $h = 1$ to 8 and $h = 1$ to 4 respectively. The winning team would be entitled to a cash reward of $500 (Australian) as long as their entry scored a MASE lower than the most accurate method in the TFC for at least one of the three data frequencies. That is, the winning entry had to score a MASE lower than 1.38 for monthly data or lower than 1.43 for quarterly data or lower than 2.28 for yearly data. The winning team would also be invited to contribute a discussion paper to the *International Journal of Forecasting* describing their methodology. All teams had access to Athanasopoulos et al. (2011) and so were aware of what methods had produced the best forecasts in the TFC.

Each team submitted their forecasts online on the Kaggle website. Once an entry was submitted it was automatically ranked on the public leaderboard. The public leaderboard was constructed by calculating the average MASE across all forecast horizons for a selected sample of 20% of the series for each frequency. As many of the series naturally clustered together (for example, a set of series representing international arrivals to Australia from various sources, or international outbound travel from Hong Kong), we selected 20% of the series from each such grouping so that the public leaderboard was constructed on a representative sample. In order to prevent teams from attempting to decode the data based on the 20% sample, we limited the daily entries to two per team. The first stage of the competition was open for submissions for 41 days and the second stage for 62 days.

In stage one of the competition, 57 teams competed. Of these, 21 teams improved on the TFC benchmark set by the Theta method of Assimakopoulos & Nikolopoulos (2000). An average of 8.6 entries per team were submitted with the top 20 teams submitting an average 15.5 entries per team. The number one ranked team was Lee C Baker who scored a MASE of 2.137 improving on the Theta benchmark by 6.65%.

In stage two of the competition there were 44 teams competing, and 11 of these improved on the TFC benchmarks set by the ARIMA algorithm of Hyndman & Khandakar (2008) for monthly data and the damped trend method for quarterly data. An average of 16.9 entries per team were submitted with the top 20 teams submitting an average 30.9 entries per team. Team Sali Mali was the winning team for the quarterly data scoring a MASE of 1.364 and improving on the benchmark by 4.74%. For the monthly data, team Lee C Baker and team Sali Mali scored a MASE of approximately 1.29 improving on the benchmark by 6.77%. This made team Sali Mali the winner of stage two of the competition and team Lee C Baker the winner of the overall competition[1]. The articles that follow describe the methods applied by the winning teams.

3 Analysis of the results and a few lessons learnt

In Table 1 we present the forecast results for the top five entries for each frequency. We consider both the MASE and the MAPE as forecast error evaluation measures. At the end of each panel we present the results for the best performing methods from the TFC.

For the monthly data all top five teams improved significantly on the average MASE calculated across all $h = 1$ to 24-steps-ahead. This improvement is less pronounced when we consider the MAPE. In this case only the most accurate of the entries, by team Sali Mali, improved on Forecast Pro and only by 1.37%. The rest of the entries were less accurate than Forecast Pro.

When forecasting monthly data, the forecast horizons of $h = 1$, $h = 12$ and $h = 24$ are of particular interest. For $h = 1$ it was only team Sali Mali that improved on the benchmarks for both the MASE and the MAPE. For the seasonal lead times of $h = 12$ and 24 there were many improvements over the benchmarks with the more pronounced improvements observed for the longer lead time.

[1] The winners donated their prize money to the Fred Hollows Foundation www.hollows.org.au

Table 1: *Forecast accuracy measures MAPE and MASE for some selected forecast horizons.*

	MAPE						MASE					
Method	Forecast horizon (h)				Average		Forecast horizon (h)				Average	
	Monthly data											
	1	2	12	24	1–12	1–24	1	2	12	24	1–12	1–24
SaliMali	15.87	16.62	20.24	22.36	18.28	19.64	0.97	1.12	1.06	1.38	1.16	1.297
LeeCBaker	16.97	19.05	21.81	23.65	18.70	20.19	1.03	1.26	1.08	1.38	1.15	1.293
Stratometrics	16.98	17.94	20.54	22.62	19.05	20.37	1.01	1.15	1.05	1.39	1.18	1.327
Robert	16.85	18.32	20.24	22.39	19.07	20.28	1.00	1.17	1.03	1.37	1.19	1.324
Idalgo	16.98	17.94	20.54	23.22	19.05	20.55	1.01	1.15	1.05	1.40	1.18	1.341
ForePro	16.75	16.22	20.54	23.27	18.38	19.91	1.02	1.17	1.12	1.54	1.22	1.401
ARIMA	17.38	17.65	21.09	24.29	19.37	21.13	1.00	1.16	1.07	1.45	1.21	1.385
	Quarterly data											
	1	2	4	8	1–4	1–8	1	2	4	8	1–4	1–8
SaliMali	11.42	11.56	13.76	21.85	12.59	14.83	1.08	1.17	1.20	1.72	1.14	1.364
LeeCBaker	12.02	12.29	14.24	21.21	13.01	15.14	1.16	1.21	1.22	1.75	1.17	1.392
Stratometrics	12.52	11.67	13.64	20.93	13.07	15.14	1.09	1.18	1.19	1.65	1.15	1.360
Robert	11.47	11.99	13.49	21.39	12.86	14.96	1.10	1.22	1.18	1.68	1.17	1.378
Idalgo	11.48	11.67	14.15	21.43	13.00	15.07	1.09	1.18	1.23	1.70	1.17	1.374
Damped	11.91	11.68	14.21	22.28	13.16	15.56	1.11	1.18	1.23	1.81	1.18	1.429
	Yearly data											
	1	2	3	4	1–2	1–4	1	2	3	4	1–2	1–4
LeeCBaker	22.80	20.12	22.31	25.69	21.46	22.73	1.34	1.87	2.43	2.91	1.60	2.137
Jtrigg	23.41	21.73	22.77	27.19	22.57	23.78	1.36	1.91	2.40	2.91	1.63	2.144
Just4fun	22.93	21.47	23.11	27.03	22.20	23.64	1.34	1.89	2.47	2.97	1.61	2.169
Germat	23.40	20.68	22.97	26.42	22.04	23.37	1.35	1.87	2.49	2.98	1.61	2.173
Strato	22.91	20.73	22.83	26.12	21.82	23.15	1.34	1.88	2.49	2.99	1.61	2.174
Theta	23.06	21.17	22.94	26.61	22.12	23.45	1.32	1.96	2.63	3.20	1.64	2.277
Naïve	21.47	20.80	24.12	28.05	21.14	23.61	1.32	2.08	2.95	3.64	1.70	2.500

The five top-ranked submissions for each data frequency from the Kaggle competitions. The benchmarks methods (shown in italics) are the best performing methods from Athanasopoulos et al. (2011)

For the quarterly data all five top teams improved significantly on the benchmark set by the damped trend method in the TFC. These improvements are observed for both the average MASE and MAPE calculated across all $h = 1$ to 8-steps-ahead. For $h = 1$ all teams improved on the benchmark for both MASE and MAPE (the only exception was team Lee C Baker for both MASE and MAPE and team Stratometrics for MAPE). Almost all teams also improved on the seasonal leads of $h = 4$ and 8 with once again the most pronounced improvements were observed for the longer lead time.

For the yearly data all five top teams improved over the Theta benchmark considering the average MASE over $h = 1$ to 4-steps-ahead. The improvements are not as pronounced when considering the average MAPE. As was also concluded in the TFC, it is extremely challenging to forecast more accurately than a random walk for $h = 1$-step ahead for yearly data. No team improved on the Naïve forecasts for $h = 1$ for either MASE or MAPE.

The top teams for the yearly data used a global trend (allowing for about 6% growth) estimated across all series. Obviously this strategy would not work in a very general forecasting competition, but where all the data come from the same industry and from a similar period of time, it proved effective. It would not have been possible to estimate the trend with any accuracy from an individual time series, but over the whole ensemble the underlying growth could estimated. However, even then, obtaining one-step forecasts that were more accurate than Naïve forecasts was not possible. We suspect that in this forecasting competition setting with immediate feedback, submissions would have improved on Naïve forecasts if the winning criterion involved some optimisation over one-step ahead forecasts for yearly data.

In previous forecasting competitions, rankings of methods have been shown to be sensitive to the accuracy measure used (Makridakis et al. 1982, Makridakis & Hibon 2000). Because we have used a public leaderboard that ranks forecasts based on a single error measure and a single forecasting task, the results are more sensitive as participants tried to improve their ranking on the public leaderboard by optimising their performance against this specific forecasting measure.

The MASE can be very sensitive to a few series, and participants found that to optimize MASE it is worth concentrating on these series. This was an unintended consequence of the measure chosen, but a similar phenomenon would have occurred with other measures including MAPE, as all scale-free measures tend to have highly skewed distributions. If possible, it would be desirable to find a metric with similar properties to MASE but with a less skewed distribution.

Some aspects of the results emphasised lessons learned from past competitions, including the fact that combining forecasts is an effective strategy for improving forecast accuracy. Other results were new. For example, leading participants found that outlier removal before forecasting can be effective, whereas methods involving outlier removal (such as Autobox) did not perform so well in the M3 competition.

The most important finding of the competition was that feedback is enormously effective in improving forecasting results. In the first part of the competition, none of the participants beat the benchmarks with their first entries, but were able to use the leaderboard feedback to learn what methods work best with the time series used. The TFC showed what off-the-shelf methods work best with these data, and the subsequent Kaggle-based competition showed how much improvement in forecast accuracy is possible with feedback-learning and then tweaking and adjusting the methods to be tuned to the particular collection of time series being forecast.

In the second part of the competition, the benchmarks were more easily beaten. Contributors found that methods developed specifically for these data were better than existing all-purpose automated time series methods. However, feedback from the leader board was still used in order to seek further improvements.

4 Conclusions

This style of competition was also used by Netflix, an American movie rental corporation in October 2006. Netflix posted a prize of US$1million in return for the algorithm that would better by 10% the accuracy of their own algorithm for predicting customers ratings of films. It took approximately four years and over 50,000 submissions for this target to be reached. Since then, various other competitions have taken place outside academia with the latest large scale competition offering a US$3million prize from the Heritage Provider Network in the US (www.heritagehealthprize.com). This competition is hosted by Kaggle, and involves predicting patient hospitalisation; it is expected to attract over 100,000 submissions.

The process of providing feedback in forecasting competitions is more in line with what happens in practice where forecasters evaluate, review and adjust their forecasting practices. Feedback provides clear motivation to forecasters to improve their practices. We suggest that it is time for the forecasting literature to move towards competitions with feedback as this is now feasible. We believe that this will make forecasting competitions simpler and quicker to conduct, which will result to more successful competitions and greater improvements in forecasting practices.

5 Acknowledgments

We thank Anthony Goldbloom, founder of Kaggle, for his assistance and hosting of the competition.

References

Assimakopoulos, V. & Nikolopoulos, K. (2000), 'The theta model: a decomposition approach to forecasting', *International Journal of Forecasting* **16**, 521–530.

Athanasopoulos, G., Hyndman, R. J., Song, H. & Wu, D. C. (2011), 'The tourism forecasting competition', *International Journal of Forecasting*, **forthcoming**, doi:10.1016/j.ijforecast.2010.04.009.

Hyndman, R. J. & Khandakar, Y. (2008), 'Automatic time series forecasting: The forecast package for R', *Journal of Statistical Software* **26**, 1–22.

Makridakis, S., Anderson, A., Carbone, R., Fildes, R., Hibon, M., Lewandowski, R., Newton, J., Parzen, E. & Winkler, R. (1982), 'The accuracy of extrapolation (time series) methods: Results of a forecasting competition', *Journal of Forecasting* **1**, 111–153.

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

Newbold, P. & Granger, C. W. J. (1974), 'Experience with forecasting univariate time series and the combination of forecasts', *Journal of the Royal Statistical Society. Series A (General)* **137**(2), 131–165.

Reid, D. (1969), A comparative study of time series prediction techniques on economic data, PhD thesis, University of Nottingham.