Framework for Comparing Accuracy of Time-Series Forecasting Methods

Junichi Sekitani
Graduate School for Creative Cities
Osaka City University
Osaka, Japan
sekitani.junichi@trans-cosmos.co.jp

Harumi Murakami
Graduate School of Informatics
Osaka Metropolitan University
Osaka, Japan
harumi@omu.ac.jp

Abstract: The research and development of time-series forecasting requires a relative assessment of forecast accuracy, although determining which model or method to select is difficult. This study creates a simple experimental framework for selecting time-series forecasting methods, based on the methods employed as benchmarks in the M4 Competition and commonly used in machine learning competitions. We added gradient boosting and other methods used in this study. Our experimental results using M4 data confirmed the high accuracy of the combination and statistical models as in the M4 Competition.

Keywords: machine learning, forecasting competitions, time-series methods, benchmarking methods

I. INTRODUCTION

Artificial intelligence has attracted much attention for its applications to such real-world problems as competing against world-class professionals in *Go* and *Shogi* (Japanese chess) and automated-automobile driving. Against this background, artificial intelligence has also been applied in the field of forecasting, and time-series forecasting with machine learning methods has been proposed as an alternative to statistical models [1]. However, in the research and development of time-series forecasting with machine learning methods, since few proposals have objectively evaluated the accuracy of forecasting [2], its relative accuracy must be evaluated by benchmarking, which was even an issue before machine learning [3]. Determining which model or method to select for solving the time-series forecasting problem is also difficult. Although such forecasting competitions as M competitions exist [4], their results might not be adopted due to bias in the methods and methodological finessing by the participants.

This study's purpose is to create a simple, standard experimental framework for the selection of time-series forecasting methods by employing a combination of statistical and machine learning models as representative methods. As benchmarks, we used the methods from the M4 Competition [5] as a basis and added gradient boosting and other methods commonly used in machine learning competitions.

This paper reports the results of an experiment that compared the accuracy of time-series forecasting methods using data from the M4 Competition.

II. METHOD

The 12 benchmarks, including two standards, of the M4 Competition (hereafter M4 Benchmarks) are biased with only nine statistical models, two machine learning models, and one combination model. We added 19 new methods, including gradient boosting, which is often used in machine learning competitions. Tables Ⅰ, II, and III show the 31 methods. The number of combinations is 12 for statistics, 13 for machine learning, and 6 for combinations.

TABLE I
31 METHODS EMPLOYED (1): STATISTICAL

Methods			M4	Description
Statistical	Naive	Naive 1	✓	Future values are assumed to be identical as last known observed value
		Naive S	✓	Assumed identical as last known observation for the same period
		Naive 2	✓	Identical as Naive 1, although seasonally adjusted by multiplicative decomposition, if necessary
	Exponential Smoothing	SES	✓	Data are extrapolated, assuming exponential smoothing, seasonal adjustment, and no trends
		Holt	✓	Data are assumed to be exponentially smoothed, seasonally adjusted, and extrapolated, assuming a linear trend.
		Damped	✓	Data are extrapolated assuming exponential smoothing and seasonally adjusted. Extrapolated assuming a decreasing trend.
		Theta	✓	Simple exponential smoothing with drift
		ETS	✓	Automatic search for optimal parameters for exponential smoothing, e.g., AIC.
		TBATS		Exponential smoothing state space with Box-Cox transformation
	ARIMA	ARMA		Autoregressive moving average
		ARIMA	✓	ARIMA for stepwise search for optimal parameters
		Auto ARIMA		Automatic estimation of SARIMA and ARIMA in parameter search

(NOTE) M4: BENCHMARK ADOPTED BY M4 COMPETITION

TABLE II
31 METHODS EMPLOYED (2): MACHINE LEARNING

Methods			M4	Description
Machine Learning	Linear Regression	Linear		Multivariate linear regression
		Ridge		Multivariate linear regression with a L2 regularization term
		Elastic-Net		Multivariate linear regression with L1 and L2 regularizations
		GAM		Linear regression allowing the expression of nonlinear relationships
	Decision Trees	Decision Tree		Created decision trees from data
		RandomForest		Ensemble of multiple weak learner decision trees in parallel
		GBDT		Gradient boosting with multiple sequential weak learner decision trees
		XGBoost		A type of GBDT: optimized distributed gradient boosting
		LightGBM		An improvement on XGBoost, gradient boosting with reduced computation, and increased learning speed
	SVM	SVM		Extended perceptron, applying margin maximization, and kernel functions
	Neural Network	MLP	✓	Three or more layers of forward propagating neural networks
		RNN	✓	A recurrent neural network that can handle time-series data
		DeepAR		RNN-based algorithms provided by Amazon SageMaker

NOTE) M4: BENCHMARK ADOPTED BY M4 COMPETITION

TABLE III
31 METHODS EMPLOYED (3): COMBINATIONS

Methods		M4	Description
Combination	Comb 1	✓	Combination of three exponential smoothings: SES, Holt, and Damped
	Comb 2		Combination of three accurate methods: Theta, TBATS, Damped
	Comb 3		Combination of three different types of methods (Tables 1 and 2, column 2) with high accuracy: Theta, Auto ARIMA, and Random Forest
	Comb 4		Combination of two accurate statistical models: Theta and TBATS
	Comb 5		Combination of two accurate machine learning models: Random Forest and GBDT
	Comb 6		Combination of two accurate statistical models and a machine learning model: Theta, Random Forest

(NOTE) M4: BENCHMARK ADOPTED BY M4 COMPETITION

We compared the accuracy of the forecasting methods on 100,000 time-series data used in the M4 Competition, ranked by OWA using the same two metrics: sMAPE and MASE:

$$OWA = \frac{\frac{sMAPE_a}{sMAPE_b} + \frac{MASE_a}{MASE_b}}{2}$$

where a is the target forecasting method and b is the baseline forecasting method. The baseline forecasting method is Naïve 2, the same as in the M4 Competition. The M4 benchmark was recalculated from data in the M4 paper [5], and methods other than the M4 benchmark were implemented and evaluated.

III. RESULTS

The accuracy of each method was output for data from six frequencies and six domains. Table IV shows the accuracy of each method in terms of frequency, domain, and overall OWA. The columns in Table IV are ranked by overall OWA. Table V shows the accuracy of sMAPE and MASE by frequency and overall, and Table VI shows the accuracy of sMAPE and MASE by domain and overall. Tables IV, V, and VI were created by referring to the "Evaluation and Ranks" file in the M4 GitHub repository (https://github.com/Mcompetitions/M4-methods). Here we report the results of the overall OWA ranking in Table IV.

The top-ranked forecasting method was Comb 3, which integrated Theta, Auto Arima, and Random Forest. As with the top result of the M4 Competition (by the participants), the combination of statistics and machine learning was the most accurate. The next most accurate method was Comb 4, which integrated Theta and TBATS, and the top four positions were occupied by combinations. That is, the combination is good, a result that is identical as that of the M4 Competition. These combinations were more accurate than Comb 1, which is the M4 benchmark. Other than the combination method, Theta (a statistical model) was the most accurate method, followed by TBATS, a statistical model newly added in our experiment.

The most accurate machine learning method was Random Forest, which was more accurate than such statistical models as ARIMA and Holt. On the other hand, GBDT had the highest accuracy in gradient boosting, which is often used in machine learning competitions, although it was lower than Naïve 2, a simple prediction method. All the newly added machine learning methods (except Random Forest) were less accurate than Naïve 2, a simple prediction method; both were more accurate than the two M4 benchmark machine learning methods: MLP and RNN.

IV. CONCLUSIONS

As in the M4 Competition, the combination and statistical models were highly accurate. Among the machine learning methods, Random Forest was the most accurate, even outperforming the M4 benchmark in several statistical models. Such additional machine learning methods as gradient boosting that are commonly used in machine learning competitions were more accurate than MLP and RNN, which are the M4 benchmarks of machine learning models. We experimentally confirmed the effectiveness of adding such prediction methods as decision trees and gradient boosting to M4 benchmarks.

Since our comparison experiment evaluated the overall accuracy of each method, future research must comprehensively investigate the frequency and domain accuracy.

REFERENCES

[1] G. Bontempi, S. B. Taieb, and Y.-A. Le Borgne, "Machine learning strategies for time series forecasting," *European business intelligence summer school*, Springer, Berlin, Heidelberg, pp. 62-77, 2012.

[2] S. Makridakis, E. Spiliotis, and V. Assimakopoulos, "Statistical and machine learning forecasting methods: concerns and ways forward," *PLoS ONE*, vol. 13, no. 3, 2018.

[3] E. Keogh and S. Kasetty, "On the need for time series data mining benchmarks: a survey and empirical demonstration," *Data Mining and knowledge discovery,* vol. 7, pp. 349-371, 2003.

[4] R. J. Hyndman, "A brief history of forecasting competitions," *International Journal of Forecasting*, vol. 36, no. 1, pp. 7-14, 2020.

[5] S. Makridakis, E. Spiliotis, and V. Assimakopoulos, "The M4 Competition: 100,000 time series and 61 forecasting methods," *International Journal of Forecasting*, vol. 36, no. 1, pp. 54-74, 2020.

TABLE IV
OWA AND RANKING

Method	OWA													Rank
	FREQUENCY						*DOMAIN*						*Total*	
	Yearly	*Quarterly*	*Monthly*	*Weekly*	*Daily*	*Hourly*	*Macro*	*Micro*	*Demographic*	*Industry*	*Finance*	*Other*		
Comb 3	0.818	0.900	0.861	0.832	0.986	1.036	0.870	0.851	0.803	0.891	0.864	1.021	0.864	1
Comb 4	0.844	0.907	0.879	0.942	0.983	0.854	0.870	0.865	0.826	0.909	0.893	0.936	0.878	2
Comb 2	0.840	0.898	0.884	0.941	0.966	0.886	0.960	0.985	0.949	0.943	0.944	0.988	0.879	3
Comb 6	0.849	0.926	0.879	0.834	1.006	1.125	0.875	0.889	0.829	0.904	0.879	0.954	0.881	4
Theta	0.872	0.933	0.890	0.982	0.993	1.016	0.895	0.860	0.831	0.920	0.885	1.084	0.886	5
Comb 1	0.894	0.913	0.926	0.961	0.965	1.262	0.890	0.864	0.854	0.936	0.911	1.027	0.896	6
TBATS	0.967	0.948	0.918	0.927	0.983	0.776	0.919	0.938	0.914	0.926	0.894	1.019	0.925	7
Damped	0.921	0.919	0.929	0.954	0.985	1.005	0.915	0.946	0.839	0.937	0.949	0.939	0.928	8
ETS	0.952	0.926	0.922	0.961	0.995	1.052	0.925	0.953	0.919	0.933	0.900	0.959	0.929	9
Auto ARIMA	1.059	0.985	0.915	0.901	0.966	0.968	0.920	0.959	0.839	0.933	0.927	1.062	0.929	10
SES	1.003	0.961	0.940	0.976	0.993	0.991	0.937	0.964	0.843	0.941	0.972	1.077	0.948	11
Random Forest	0.925	1.002	0.941	0.826	1.028	1.340	0.970	0.945	0.897	0.977	0.935	1.244	0.955	12
Comb 5	0.934	1.006	0.958	0.824	1.031	1.988	0.988	0.953	0.899	0.991	0.952	1.648	0.981	13
ARIMA	0.900	0.975	0.970	0.991	1.002	3.365	0.940	0.975	0.928	1.017	0.936	2.156	0.993	14
Naive2	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	15
Holt	0.980	0.971	1.012	1.015	0.984	2.410	1.004	1.036	0.974	0.998	0.962	1.428	1.017	16
GBDT	0.965	1.039	0.992	0.879	1.056	2.588	1.020	0.981	0.937	1.026	0.988	2.019	1.025	17
Linear	1.011	1.073	0.997	1.039	1.296	3.521	1.021	0.983	0.958	1.039	1.028	2.681	1.060	18
XGBoost	0.995	1.397	1.003	0.897	1.346	2.755	1.046	1.069	1.040	1.053	1.086	1.682	1.075	19
Naïve 1	1.000	1.054	1.064	1.000	1.000	2.830	1.043	1.043	1.084	1.076	1.031	1.526	1.076	20
GAM	1.211	1.108	1.038	1.056	1.336	2.694	1.068	1.131	1.064	1.059	1.059	1.774	1.092	21
Naïve S	1.000	1.138	1.113	1.000	1.000	0.861	1.107	1.054	1.103	1.097	1.110	0.986	1.093	22
ARMA	1.118	1.112	1.138	0.969	1.032	9.370	1.157	1.209	1.087	1.118	1.067	4.081	1.207	23
SVM	1.255	1.318	1.131	1.317	2.300	4.619	1.181	1.197	1.187	1.239	1.161	2.437	1.219	24
Decision Tree	1.212	1.375	1.207	1.216	1.333	1.230	1.238	1.245	1.233	1.214	1.237	1.328	1.230	25
LightGBM	0.974	1.379	0.997	0.908	1.178	15.184	1.141	1.068	0.936	1.059	0.991	7.340	1.233	26
Elastic-Net	2.106	1.747	1.035	2.284	1.591	2.584	1.321	1.384	1.257	1.247	1.268	2.036	1.317	27
DeepAR	1.707	1.525	1.223	1.326	1.677	1.401	1.328	1.401	1.289	1.294	1.356	1.509	1.342	28
Ridge	1.832	1.705	1.232	1.901	1.353	2.575	1.547	1.599	1.452	1.393	1.392	1.656	1.371	29
RNN	1.328	1.524	1.537	1.647	1.891	1.542	1.927	1.348	1.100	1.174	1.253	1.963	1.404	30
MLP	1.291	1.654	1.663	2.619	3.195	0.896	1.601	1.683	1.652	1.445	1.519	2.452	1.504	31

TABLE V
ACCURACY BY FREQUENCY: SMAPE AND MASE

Method	sMAPE							MASE						
	Yearly	*Quarterly*	*Monthly*	*Weekly*	*Daily*	*Hourly*	*Total*	*Yearly*	*Quarterly*	*Monthly*	*Weekly*	*Daily*	*Hourly*	*Total*
Comb 3	13.324	9.978	12.534	7.427	3.030	20.197	11.895	3.063	1.218	1.180	1.739	2.928	1.293	2.013
Comb 4	13.825	10.070	12.629	8.648	3.020	14.926	11.950	3.148	1.226	1.219	1.915	2.918	1.189	2.066
Comb 2	13.806	9.988	12.755	8.669	2.959	16.137	12.036	3.122	1.214	1.223	1.909	2.877	1.188	2.064
Comb 6	13.853	10.269	12.784	7.375	3.092	22.638	12.206	3.176	1.255	1.205	1.759	2.986	1.351	2.063
Theta	14.357	10.341	12.804	9.104	3.055	18.138	12.219	3.233	1.263	1.233	1.976	2.944	1.388	2.103
Comb 1	14.848	10.175	13.434	8.944	2.980	22.053	12.730	3.289	1.231	1.274	1.930	2.851	1.758	2.152
TBATS	15.432	10.523	13.179	8.430	3.008	13.319	12.537	3.700	1.282	1.276	1.903	2.935	1.099	2.201
Damped	15.198	10.237	13.473	8.866	3.064	19.265	12.764	3.407	1.240	1.279	1.918	2.889	1.277	2.165
ETS	15.845	10.255	13.274	8.771	3.057	18.417	12.688	3.493	1.255	1.276	1.966	2.953	1.463	2.179
Auto ARIMA	16.359	10.763	13.156	8.240	2.941	17.164	12.714	4.169	1.353	1.269	1.838	2.895	1.331	2.264
SES	16.396	10.600	13.618	9.012	3.045	18.094	13.027	3.744	1.308	1.294	1.973	2.955	1.325	2.237
Random Forest	15.300	11.179	13.866	7.145	3.154	29.177	13.335	3.410	1.348	1.274	1.777	3.059	1.452	2.191
Comb 5	15.480	11.189	14.138	7.222	3.185	49.307	13.854	3.437	1.359	1.294	1.755	3.046	1.717	2.224
ARIMA	15.119	10.922	14.149	9.237	3.144	53.244	13.849	3.266	1.307	1.326	1.983	2.915	5.090	2.280
Naive2	16.342	11.012	14.427	9.161	3.045	18.383	13.636	3.736	1.364	1.383	2.039	2.998	1.327	2.350
Holt	16.354	10.907	14.812	9.708	3.066	29.249	14.054	3.581	1.297	1.378	1.978	2.884	4.286	2.356
GBDT	16.058	11.533	14.644	7.875	3.299	65.292	14.566	3.541	1.406	1.340	1.832	3.084	2.157	2.306
Linear	16.981	12.030	14.698	10.053	4.034	95.291	15.286	3.668	1.437	1.349	2.000	3.796	2.466	2.349
XGBoost	16.110	15.182	14.503	7.611	4.046	21.020	14.371	3.753	1.930	1.383	1.965	4.086	5.796	2.576
Naïve 1	16.342	11.610	15.257	9.161	3.045	43.003	14.670	3.736	1.438	1.481	2.039	2.998	4.408	2.527
GAM	19.891	12.090	15.027	8.539	4.466	20.618	14.685	4.501	1.525	1.431	2.407	3.613	5.663	2.601
Naïve S	16.342	12.521	15.988	9.161	3.045	13.912	14.849	3.736	1.554	1.545	2.039	2.998	1.281	2.579
ARMA	18.523	12.434	16.504	8.663	3.185	34.398	15.743	4.119	1.493	1.566	2.022	3.050	22.390	2.961
SVM	20.144	13.940	15.801	9.588	8.191	19.985	15.680	4.773	1.868	1.613	3.236	5.724	10.819	3.026
Decision Tree	19.771	15.024	17.161	10.573	4.104	18.272	16.510	4.534	1.890	1.692	2.605	3.952	1.947	2.935
LightGBM	15.723	14.875	14.523	7.667	3.625	197.361	17.019	3.686	1.919	1.366	1.995	3.496	26.057	2.860
Elastic-Net	33.839	18.356	14.776	15.820	5.141	27.937	17.135	8.000	2.493	1.445	5.791	4.477	4.843	3.237
DeepAR	26.783	16.775	17.604	11.070	4.963	17.276	17.853	6.628	2.083	1.695	2.944	5.168	2.471	3.230
Ridge	29.208	17.866	18.091	13.435	4.373	27.879	18.758	7.009	2.439	1.673	4.763	3.807	4.824	3.364
RNN	22.398	17.027	24.056	15.220	5.964	14.698	21.805	4.801	2.049	1.944	3.330	5.469	3.031	3.321
MLP	21.764	18.500	24.333	21.349	9.321	13.842	22.310	4.668	2.221	2.267	5.926	9.977	1.380	3.739

TABLE VI
ACCURACY BY DOMAIN: SMAPE AND MASE

Method	sMAPE							MASE						
	Yearly	*Quarterly*	*Monthly*	*Weekly*	*Daily*	*Hourly*	*Total*	*Yearly*	*Quarterly*	*Monthly*	*Weekly*	*Daily*	*Hourly*	*Total*
Comb 3	12.703	12.108	5.524	12.756	12.724	13.286	11.895	1.247	2.531	1.132	1.415	3.025	1.763	2.013
Comb 4	12.632	12.214	5.687	12.969	12.990	11.259	11.950	1.254	2.594	1.164	1.448	3.167	1.769	2.066
Comb 2	13.867	13.581	7.047	13.451	14.057	11.472	13.027	1.390	3.027	1.245	1.501	3.270	1.938	1.885
Comb 6	12.736	12.603	5.706	12.892	12.852	11.734	12.036	1.256	2.658	1.167	1.440	3.105	1.759	2.064
Theta	13.045	12.190	5.720	13.129	13.110	14.408	12.206	1.284	2.569	1.171	1.464	3.086	1.817	2.063
Comb 1	12.830	12.128	5.869	13.333	13.441	12.748	12.219	1.290	2.609	1.204	1.492	3.184	1.876	2.103
TBATS	13.382	12.840	6.774	13.264	13.272	11.444	12.730	1.320	2.906	1.200	1.468	3.106	2.064	2.152
Damped	13.311	13.340	5.770	13.343	13.486	10.897	12.537	1.317	2.841	1.182	1.496	3.441	1.842	2.201
ETS	13.507	13.058	6.821	13.370	13.308	11.275	12.764	1.325	2.947	1.207	1.481	3.140	1.859	2.165
Auto ARIMA	13.377	13.500	5.788	13.253	13.634	13.021	12.688	1.322	2.887	1.178	1.492	3.254	1.966	2.179
SES	13.365	13.407	5.767	13.308	13.846	12.792	12.714	1.373	2.935	1.194	1.513	3.514	2.066	2.264
Random Forest	14.324	13.607	6.203	14.010	13.973	17.519	13.335	1.371	2.779	1.258	1.546	3.226	1.920	2.191
Comb 5	14.616	13.674	6.225	14.282	14.279	25.996	13.854	1.394	2.812	1.259	1.563	3.274	2.071	2.224
ARIMA	13.780	13.941	6.441	14.648	14.003	27.781	13.849	1.339	2.884	1.297	1.604	3.231	3.766	2.280
Naive2	14.278	13.741	7.403	14.205	14.789	11.605	13.636	1.465	3.085	1.316	1.600	3.489	1.963	2.350
Holt	14.746	14.298	7.282	14.441	14.356	13.117	14.054	1.429	3.184	1.268	1.566	3.325	3.388	2.356
GBDT	15.098	14.003	6.535	14.841	14.863	32.962	14.566	1.440	2.912	1.303	1.611	3.392	2.349	2.306
Linear	15.019	13.993	6.641	14.986	15.654	46.513	15.286	1.451	2.926	1.340	1.636	3.479	2.659	2.349
XGBoost	14.992	14.935	7.191	14.823	15.872	14.426	14.371	1.528	3.242	1.460	1.701	3.832	4.162	2.576
Naïve 1	14.791	14.219	7.949	15.072	15.223	14.719	14.670	1.539	3.244	1.441	1.746	3.601	3.502	2.527
GAM	15.393	15.748	7.259	15.006	15.746	15.210	14.685	1.549	3.441	1.510	1.698	3.677	4.390	2.601
Naïve S	15.677	14.490	8.085	15.420	16.249	11.240	14.849	1.634	3.253	1.466	1.772	3.909	1.972	2.579
ARMA	17.000	17.134	7.554	15.858	15.801	20.688	15.743	1.647	3.614	1.518	1.791	3.718	12.524	2.961
SVM	16.635	16.229	7.947	16.715	16.740	15.364	15.680	1.755	3.742	1.713	2.082	4.151	6.968	3.026
Decision Tree	17.490	17.143	8.508	17.052	18.166	14.996	16.510	1.832	3.835	1.733	1.965	4.345	2.677	2.935
LightGBM	16.303	14.993	6.419	14.862	14.648	87.001	17.019	1.671	3.221	1.321	1.716	3.460	14.101	2.860
Elastic-Net	18.408	18.223	8.423	17.184	17.993	20.574	17.135	1.981	4.446	1.810	2.053	4.606	4.513	3.237
DeepAR	18.765	19.312	8.881	18.026	19.580	15.355	17.853	1.966	4.307	1.814	2.110	4.845	3.327	3.230
Ridge	23.471	24.286	10.942	20.181	20.907	18.093	21.805	2.124	4.417	1.877	2.184	4.781	3.440	2.685
RNN	27.980	18.150	7.560	16.455	18.122	19.840	18.758	2.775	4.240	1.551	1.904	4.467	4.350	3.364
MLP	23.279	23.947	12.218	20.897	22.296	19.173	22.310	2.301	5.009	2.176	2.271	5.339	6.382	3.225