A Fast and Scalable Ensemble of Global Models with Long Memory and Data Partitioning for the M5 forecasting competition

2022 International Symposium on Forecasting
July 13, 2022

Presenter

Kasun Bandara

Teamed with

Hansika Hewamalage
Rakshitha Godahewa
Puwasala Gamakumara

EnergyAustralia

University of Melbourne

Outline

M5 Competition Overview

- The fifth round of the famous M competitions
- Accurately forecast 42,840 hierarchically organised time series representing the sales demand of 3049 products sold by Walmart.
  - Required to submit 30,490 point forecasts for the lowest level of the hierarchy (store-product combinations)
  - Prediction horizon of 28 days
  - Validation phase: Allowing the teams to fine tune the model performance, Test phase: Used to evaluate the final performance of the teams.
- Two submission tracks: 1) Accuracy 2) Uncertainty

M5 time series structure

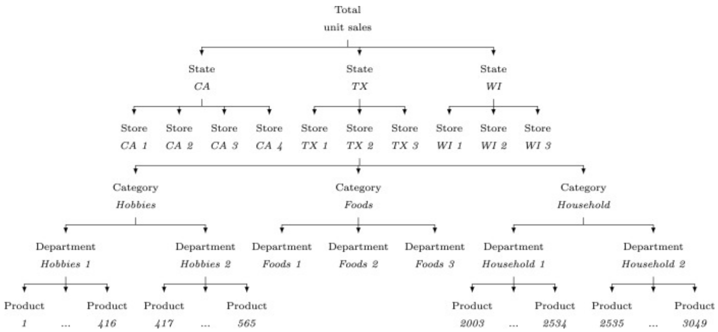

Figure: Time series hierarchy used in the M5 competition [Makridakis et al., 2021]

Time series aggregation structure

Level id	Level Description	Aggregation Level	Number of series
1	Unit sales of all products, aggregated for all stores/states	Total	1
2	Unit sales of all products, aggregated for each State	State	3
3	Unit sales of all products, aggregated for each store	Store	10
4	Unit sales of all products, aggregated for each category	Category	3
5	Unit sales of all products, aggregated for each department	Department	7
6	Unit sales of all products, aggregated for each State and category	State-Category	9
7	Unit sales of all products, aggregated for each State and department	State-Department	21
8	Unit sales of all products, aggregated for each store and category	Store-Category	30
9	Unit sales of all products, aggregated for each store and department	Store-Department	70
10	Unit sales of product x, aggregated for all stores/states	Product	3,049
11	Unit sales of product x, aggregated for each State	Product-State	9,147
12	Unit sales of product x, aggregated for each store	Product-Store	30,490
	Total		**42,840**

Figure: Aggregation levels of the M5 competition [Makridakis et al., 2021]

Problem Motivation

- Generating accurate and reliable retail demand forecasts is an important endeavour for the retailers.
  - Better demand planning and efficient resource management.
  - Forecasts at different levels of aggregation can be important for certain managerial decisions
- Vital for supermarkets as it provides better grounds for decision-making of organisational short-term, medium-term and long-term goals.
  - Accurate forecasts offer significant savings and cost reductions.
  - Overestimation or Underestimation leads to product spillovers and unmet customer demand.

Competition Challenges

- The business environment in retail is highly dynamic and often volatile, which is largely caused by holiday effects, low product-sales conversion rate, competitor behaviour
  - highly non-stationary historical data
  - irregular sales patterns
  - influence of exogenous variables
  - hierarchically organised large collection of time series

Evaluation

- Error measures
  - Weighted root mean squared scaled error: WRMSSE

$$\text{RMSSE} = \sqrt{\frac{\frac{1}{h}\sum_{t=n+1}^{n+h}\left(y_t - \widehat{y_t}\right)^2}{\frac{1}{n-1}\sum_{t=2}^{n}\left(y_t - y_{t-1}\right)^2}}$$

$$\text{WRMSSE} = \sum_{i=1}^{42840} w_i * RMSSE_i$$

y_t: actual observation at time t

$\widehat{y_t}$: Forecast at time t

h: Number of data points in the test set (forecast horizon)

n: Number of data points in the training set

Proposed Solution

- A four-layered cross-learning-based retail demand forecast framework.
- Achieved 17th position in the accuracy track (Top 1%)
- Pre-processing layer, a Model prediction layer, a Post-processing layer, and an Ensembling layer
  - Pre-processing: time series grouping, normalisation
  - Model prediction: application of global models.
  - Post-processing: data denormalisation
  - Ensembling: model combination strategy.

Retail Demand Forecasting Framework

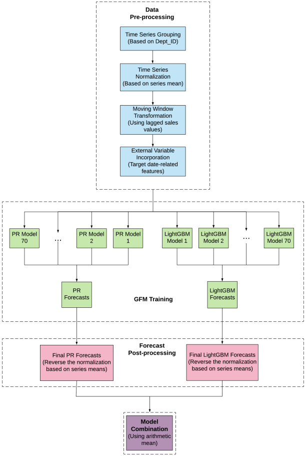

Figure: The overall summary of the proposed retail demand forecasting framework

Outline

Global Forecast Models (GFMs)

- Methods that estimate model parameters jointly from all available time series [Januschowski et al., 2020].
- A unified forecasting model that is built using a collection of time series
  - Borrow similar behaviours and structures from other related time series.
  - Improves model generalizability.
  - Adequate data for model fitting.
  - Ability to exploit the cross-series information.
- Forecasting a large quantity of related time series: "Related" in terms of similarity of their DGP (not necessarily mere correlations)

Complexity of GFMs

- Global models can afford to be more complex.
- Complexity can be added as:
  - Longer memory (longer input windows, more lags)
  - Non-linear/non-parametric models (NN variants, GBT, . . .)
  - Data partitioning (Time series clustering)
- GFMs can be designed with a much higher complexity, yet still achieve better generalisation error than the univariate models for larger datasets [Montero-Manso and Hyndman, 2021]

GFMs for Forecast combination

- Ensembling works in forecasting just as well as in any other area.
- Local models to capture unique behaviours, Global models to capture common patterns.
- Ensembles of local and global models
  - Energy Demand Forecasting [Triguero, 2020]
  - Weekly time series forecasting [Godahewa et al., 2021]

Date Pre-processing Layer

- Time series grouping
  - time series grouping at the department level (**data partitioning**)
- Data normalisation
  - Using the *meanscale* transformation strategy to account for sales scale differences
- Feature engineering
  - *day-of-week, day-of-month, month, is-working-day, is-weekend, snap, and events* as temporal features
  - 400 days of sales lags (**longer memory**)

Forecast Engine

- Light Gradient Boosting Machine (LightGBM) and Pooled Regression (PR) as the main prediction models
  - Both trained globally across a collection of time series using exogenous variables
  - The recursive strategy to generate multi step-ahead forecasts
- Loss functions
  - Poisson loss as the loss function of LightGBM.
  - Re-weighted Least-Squares as the loss function of PR models
- Hyperparameter selection and optimization
  - The last 28 days prior to the forecast horizon and the same 28 days as the forecast horizon from the last year used as the validation set.
  - A grid-based methodology to minimise the WRMSSE error measure

LightGBM model [Ke et al., 2017]

- A popular and computationally efficient machine learning algorithm
- A variant of Gradient Boosting Models (GBM)
  - Combines many weak learners to come up with one strong learner
  - The initial "weak" decision tree is "boosted" to produce more accurate forecasts
  - Used the implementation available from the R package *lightgbm*

LightGBM Hyperparameter grid

Table: Parameter value ranges used to train the LightGBM models in our experiments

Parameter	Min. value	Max. value	Opt. Value
Bagging frequency	1	5	1
Bagging fraction	0.25	1	0.75
L2 regularization	0	0.5	0.1
Learning rate	0.025	0.1	0.075
Number of leaves	30	320	128
Min no inst per leaf	50	150	100
No of iterations	500	1500	1200

Pooled Regression model

- A global version of an Auto-Regression (AR) model of order 400 (lags of sales) along with the external variables
- The term pooling indicates that one model is built using many series
- Using the implementation available in the **glm** function from the R package *glmnet*

$$y_t = c + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \phi_3 y_{t-3} + \ldots + \phi_p y_{t-p} + \sum_{i=1}^{n} \alpha_i d_i + \sum_{i=1}^{m} \beta_i e_i + \epsilon_t$$

y_{t-1} to y_{t-p} denote the lags of sales
ϕ_1 to ϕ_p indicate the coefficients of the model
d_i: data related variables
e_i: events variables

Post-processing Layer

- Reversing the initial preprocessing steps.
  - Multiplying the forecasts by the mean of the respective series (For global forecasting models).
- The sales forecasts generated for each department id is collated (forecasts generated for each cluster)

Ensembling Layer

- In time series forecasting, ensemble models are mostly used in the form of forecast combinations.
- The model diversity to the forecast combination was introduced by employing both linear (PR) and nonlinear (LightGBM) global models in our forecast framework[Lichtendahl and Winkler, 2020]
  - Compute the simple average of PR and LightGBM model forecast.

Outline

1 Introduction

2 Design and Development

3 Evaluation

4 Winning solution Vs Ours

Model Results

Rank	Team	Aggregation level 1 Total	2 State	3 Store	4 Category	5 Department	6 State Category	7 State Department	8 Store Category	9 Store Department	10 Product	11 Product State	12 Product Store	Average	Improvement over ES_bu (%)
1	YJ_STU	0.199	0.310	**0.400**	0.277	0.365	0.390	**0.474**	**0.480**	**0.573**	0.966	0.929	0.884	**0.520**	22.4
2	Matthias	0.186	**0.294**	0.416	**0.246**	**0.349**	**0.381**	0.481	0.497	0.594	1.023	0.964	0.907	0.528	21.3
3	mf	0.236	0.319	0.421	0.308	0.397	0.405	0.496	0.505	0.600	**0.950**	**0.917**	0.875	0.536	20.2
4	monsaraida	0.254	0.340	0.418	0.302	0.377	0.411	0.483	0.490	0.579	0.963	0.928	0.886	0.536	20.1
5	Alan Lahoud	0.213	0.324	0.414	0.272	0.361	0.416	0.494	0.503	0.595	0.995	0.950	0.897	0.536	20.1
6	wyzJack_STU	0.248	0.367	0.431	0.319	0.396	0.436	0.502	0.502	0.584	0.953	0.918	**0.875**	0.544	18.9
7	RandomLearner	0.194	0.317	0.423	0.276	0.404	0.408	0.516	0.503	0.608	1.029	0.968	0.910	0.546	18.6
8	SHJ	0.279	0.357	0.419	0.336	0.406	0.429	0.498	0.497	0.586	0.956	0.922	0.878	0.547	18.5
9	gest 2	0.197	0.322	0.424	0.269	0.406	0.420	0.536	0.513	0.624	1.000	0.953	0.901	0.547	18.5
10	DenisKokosinskiy_STU	0.294	0.363	0.419	0.341	0.401	0.429	0.493	0.494	0.581	0.955	0.921	0.878	0.547	18.4
11	XueWang	0.288	0.358	0.417	0.348	0.423	0.424	0.501	0.490	0.582	0.959	0.921	0.876	0.549	18.2
12	yq_STU	0.226	0.320	0.456	0.294	0.403	0.399	0.496	0.526	0.614	1.010	0.954	0.899	0.550	18.1
13	PoHaoChou	0.212	0.317	0.459	0.322	0.402	0.413	0.504	0.539	0.630	0.968	0.940	0.898	0.550	18.0
14	Tsuru	0.257	0.335	0.402	0.325	0.416	0.421	0.506	0.503	0.608	0.994	0.951	0.900	0.552	17.8
15	bk_18	0.217	0.333	0.420	0.303	0.433	0.431	0.537	0.510	0.615	0.986	0.943	0.893	0.552	17.8
16	N60610	0.195	0.350	0.436	0.298	0.409	0.441	0.530	0.521	0.619	0.976	0.945	0.900	0.552	17.8
17	MonashSL_STU	0.247	0.342	0.446	0.308	0.404	0.412	0.501	0.520	0.622	0.992	0.944	0.892	0.552	17.7
18	leoclement	0.270	0.354	0.410	0.322	0.415	0.434	0.526	0.498	0.603	0.986	0.945	0.896	0.555	17.3
19	minghui_Tju	0.254	0.365	0.428	0.327	0.425	0.439	0.529	0.505	0.602	0.979	0.936	0.886	0.556	17.1
20	zfc613	0.236	0.343	0.470	0.291	0.387	0.412	0.506	0.539	0.627	1.008	0.959	0.907	0.557	17.0
21	Nodalpoints	0.312	0.376	0.423	0.353	0.416	0.443	0.508	0.500	0.587	0.963	0.925	0.880	0.557	17.0
22	CPUkiller	0.263	0.367	0.427	0.333	0.431	0.438	0.529	0.504	0.601	0.979	0.935	0.885	0.558	16.9
23	dont overfit	0.263	0.367	0.427	0.333	0.431	0.438	0.529	0.504	0.601	0.979	0.935	0.885	0.558	16.9
24	Dan Hargreaves	0.269	0.350	0.439	0.316	0.405	0.431	0.517	0.522	0.616	0.984	0.946	0.900	0.558	16.9
25	M0T0_STU	0.252	0.346	0.425	0.345	0.431	0.445	0.530	0.526	0.623	0.967	0.932	0.886	0.559	16.7
26	Genryu	0.295	0.368	0.436	0.340	0.410	0.445	0.515	0.523	0.611	0.961	0.924	0.880	0.559	16.7
27	Moscow Five	0.245	0.349	0.442	0.309	0.435	0.436	0.542	0.526	0.624	0.988	0.941	0.889	0.560	16.5
28	Daniela A	**0.162**	0.333	0.483	0.278	0.441	0.412	0.569	0.558	0.679	0.988	0.942	0.892	0.561	16.3
29	shuheioka	0.267	0.354	0.440	0.313	0.419	0.430	0.524	0.517	0.616	1.002	0.954	0.903	0.562	16.3
30	sk 2	0.191	0.381	0.511	0.263	0.364	0.470	0.552	0.585	0.661	0.962	0.932	0.887	0.563	16.1
31	nagao	0.279	0.382	0.443	0.328	0.413	0.456	0.531	0.523	0.609	0.975	0.936	0.889	0.564	16.0
32	AjayNagar	0.221	0.324	0.518	0.285	0.412	0.400	0.515	0.573	0.660	1.010	0.954	0.900	0.564	15.9
33	cjwh	0.248	0.348	0.449	0.314	0.420	0.442	0.544	0.535	0.639	1.002	0.955	0.903	0.566	15.6
34	CWD75	0.237	0.326	0.422	0.330	0.452	0.442	0.551	0.526	0.637	1.004	0.960	0.912	0.567	15.6
35	Groot	0.278	0.384	0.443	0.342	0.432	0.458	0.540	0.519	0.611	0.979	0.937	0.887	0.567	15.4
36	Astral	0.299	0.381	0.453	0.342	0.401	0.453	0.520	0.528	0.611	0.984	0.945	0.896	0.568	15.4
37	Logistic	0.278	0.386	0.445	0.344	0.436	0.457	0.541	0.518	0.610	0.979	0.936	0.886	0.568	15.3
38	jdsc_perceiving_team	0.262	0.372	0.461	0.326	0.433	0.445	0.532	0.531	0.623	0.990	0.948	0.897	0.568	15.3
39	Abzal	0.314	0.373	0.434	0.351	0.420	0.447	0.519	0.515	0.603	0.998	0.956	0.906	0.570	15.1
40	Pianus	0.287	0.383	0.473	0.342	0.435	0.451	0.535	0.536	0.626	0.964	0.926	0.880	0.570	15.1
41	NAU	0.277	0.366	0.456	0.310	0.425	0.440	0.537	0.532	0.633	1.002	0.957	0.906	0.570	15.0
42	shirokane_friends	0.300	0.387	0.454	0.347	0.429	0.461	0.540	0.534	0.619	0.965	0.926	0.880	0.570	15.0
43	Alexnet	0.301	0.390	0.444	0.353	0.435	0.463	0.540	0.520	0.610	0.975	0.934	0.885	0.571	14.9
44	Griffin_Series	0.317	0.380	0.469	0.361	0.442	0.448	0.527	0.529	0.618	0.971	0.933	0.887	0.574	14.5
45	Hiromitsu Kigure	0.291	0.380	0.462	0.342	0.428	0.449	0.533	0.535	0.629	0.991	0.950	0.895	0.574	14.5
46	YK	0.247	0.369	0.464	0.314	0.438	0.453	0.551	0.542	0.644	1.011	0.958	0.904	0.575	14.4
47	PASSTA	0.339	0.396	0.460	0.366	0.421	0.457	0.521	0.532	0.614	0.970	0.933	0.886	0.575	14.4
48	golubyatniks	0.359	0.413	0.455	0.387	0.434	0.466	0.519	0.521	0.600	0.956	0.922	0.879	0.576	14.2
49	belkasanek	0.184	0.329	0.538	0.260	0.427	0.416	0.549	0.608	0.701	1.028	0.964	0.905	0.576	14.2
50	Random_prediction	0.249	0.348	0.455	0.347	0.457	0.460	0.563	0.558	0.655	0.986	0.943	0.890	0.576	14.2

Figure: Final model rankings [Makridakis et al., 2022]

Model Results Decomposition

Level	Ensemble	LightGBM	PR
1	0.247	0.340	**0.216**
2	**0.342**	0.397	0.350
3	**0.446**	0.505	0.460
4	0.308	0.384	**0.291**
5	**0.404**	0.482	**0.404**
6	**0.412**	0.458	0.423
7	**0.501**	0.558	0.524
8	**0.520**	0.567	0.539
9	**0.622**	0.684	0.643
10	0.992	1.049	**0.985**
11	0.944	0.983	**0.941**
12	0.892	0.921	**0.890**

Table: The WRMSSE values for the base models.

Outline

1 Introduction

2 Design and Development

3 Evaluation

4 Winning solution Vs Ours

Benchmark against the Winning Solution

Methodology	M5 Winning Method	Proposed method
Data Pre-processing	• At store level (10 groups), store-category level (30 groups) and department level (70 groups)	• Only at the department level (70 groups) • Mean normalisation to account for sales scale differences
Feature Engineering	• State_id, Store_id, Cat_id, Dept_id and Prod_id • Hand-crafted price features such as price, price_norm, price_max, price_min, and price_mean • day, month, year, day_of_week, week_num, month_week, is_workingday, is_weekend, snap and events as date related features • Two weeks of historical sales and sales mean, standard deviations at the store and state level for the entire training period • Sales mean and the sales standard deviation for different window sizes of one week, two weeks, one month, two months and half a year	• day_of_week, day_of_month, month, is_workingday, is_weekend, snap and events as date related features • 400 days of sales lags
Forecasting Setup	• LightGBM as the prediction model • Tweedie loss as the loss function • 9 LightGBM hyperparameters (see Appendix A) • Both recursive and direct strategies to generate multi step-ahead forecasts	• A combination of LightGBM and PR as the prediction models • Poisson loss as the loss function of LightGBM models and Re-weighted Least-Squares as the loss function of PR models • 7 LightGBM hyperparameters (see Appendix A) • The recursive strategy to generate multi step-ahead forecasts
Validation splits	• 13 validation splits corresponding to the last 13 28-day periods constructed through the rolling-origin mechanism	• Last 28 days prior to the forecast horizon and the same 28 days as the forecast horizon from the previous year

Figure: Comparison of the proposed method against the M5 winning solution [Bandara et al., 2021]

Publication

- Bandara, K. et al. 2021. A fast and scalable ensemble of global models with long memory and data partitioning for the M5 forecasting competition. International journal of forecasting. (Dec. 2021). DOI:https://doi.org/10.1016/j.ijforecast.2021.11.004.

Our Team

Kasun Bandara

Hansika Hewamalage

Rakshitha Godahewa

Puwasala Gamakumara

Thank you

Kasun Bandara

Kasun.Bandara@unimelb.edu.au

References I

Bandara, K., Hewamalage, H., Godahewa, R., and Gamakumara, P. (2021).
A fast and scalable ensemble of global models with long memory and data partitioning for the M5 forecasting competition.
Int. J. Forecast.

Godahewa, R., Bandara, K., Webb, G. I., Smyl, S., and Bergmeir, C. (2021).
Ensembles of localised models for time series forecasting.
Knowledge-Based Systems, 233:107518.

Januschowski, T., Gasthaus, J., Wang, Y., Salinas, D., Flunkert, V., Bohlke-Schneider, M., and Callot, L. (2020).
Criteria for classifying forecasting methods.
International Journal of Forecasting, 36(1):167–177.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., and Liu, T. (2017).
Lightgbm: a highly efficient gradient boosting decision tree.
In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, page 3149–3157, Red Hook, NY, USA. Curran Associates Inc.

References II

Lichtendahl, K. and Winkler, R. (2020).
Why do some combinations perform better than others?
International Journal of Forecasting, 36(1):142–149.

Makridakis, S., Spiliotis, E., and Assimakopoulos, V. (2021).
The M5 competition: Background, organization, and implementation.
Int. J. Forecast.

Makridakis, S., Spiliotis, E., and Assimakopoulos, V. (2022).
M5 accuracy competition: Results, findings, and conclusions.
Int. J. Forecast.

Montero-Manso, P. and Hyndman, R. J. (2021).
Principles and algorithms for forecasting groups of time series: Locality and globality.
International Journal of Forecasting.

Triguero, I. (2020).
Ieee-cis technical challenge on energy prediction from smart meter data.