Forecasting and predictive analytics

PROJECT

Forecasting Walmart Sales

Guillaume Chevillon (chevillon@essec.edu), Pierre Jacob (jacob@essec.edu)

Fall 2021

This project aims at forecasting sales of some products/items at three Walmart stores in California. It is based on the M5 competition that ran in 2020. M-Competitions are open-entry competitions run (inter alia) by Prof. Spyros Makridakis, of the University of Nicosia (Cyprus) and formerly of INSEAD, where teams compete in producing forecasts for thousands of time series data. In the latest two competitions, called M4 and M5, methods based on Machine Learning and Neural Networks have started to perform well relative to traditional methods. The generality of these findings has not yet been established, although the main two findings that seem to emerge is that SARIMA/Exponential Smoothing methods seem to play a role either (i) for initializing machine/deep learning techniques, or (ii) by being combined with machine/deep learning techniques. Another key finding seems to be that univariate methods are dominated overall by their multivariate counterparts, but that the latter are not yet fully developed at the intersection of machine/deep learning and the time series forecasting.

Here, you are going to perform some work based on the latest M5 Kaggle competition, but slightly adapted. The data consists of 6 different item categories (items related to hobbies, two household categories and three food categories), in three Walmart Stores labeled CA_1, CA_2 and CA_3. The data is at the daily frequency but forecasts will be evaluated both at the daily and weekly frequencies. The data are presented in Figure 4.0.1.

You will find below a set of questions, but you should feel free to refer to the M5 competition on Kaggle, as well to related articles, code and datasets, if you wish more information. You will find the M5 guidelines in the Moodle (and on github https://github.com/Mcompetitions/M5-methods). The data have been aggregated so they do not correspond to individual items so you need not consider the hierarchical methods or the weighting schemes for forecast evaluation.

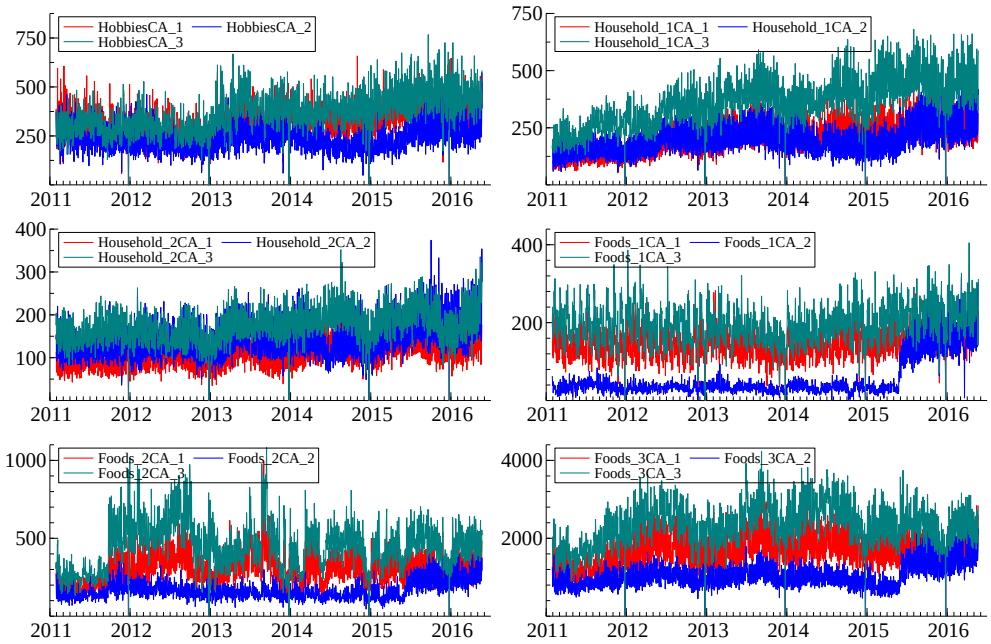

Figure 0.0.1: Daily Walmart sales

OUTPUT Your output need be some code that runs on R. Please use a Markdown for the output so we can see your code and report. Some machine learning techniques can be implemented in Python but you need to provide the corresponding code, store the forecasts and import them in R for the assessment and comparison. Make sure to store all forecasts so you can progressively add any of them in the comparisons.

DUE DATE: JANUARY 15, 2022 But feel free to hand it in beforehand.

1 Assessing the data and univariate benchmarks

Throughout your analysis, you need to split your data using a training/testing separation. You must ensure that your results are not specific to your choice of split, so you must therefore consider several such splits. We do not repeat it, so you must keep it in mind for every question.

1. Are these six variables stationary,
   (a) using a simple method, e.g. looking at the time series plots and at the ACF/PACF?
   (b) performing a test?
2. We now want to assess the set of benchmark in the M5 guidelines. For several sample splits of your choice,
   (a) Fit Statistical benchmarks 1 (Naive), 2 (sNaive), 3 (ES), 4 (MA), 13 (ESX), as well as SARIMA, SARIMAX (using X as in ESX) and Holt-Winters up to the end of the training subsample.
   (b) Consider also a state-space model where the latent state variable follows a random walk. Have a look as well at the TBATS package in R.
   (c) Obtain a sequence of 1-step ahead point forecasts over the testing subsample.
   (d) Assess the quality of the forecasts using relevant loss functions.
   (e) Compare the above with combinations of the forecasting techniques (using simple averages across methods, such as Combination benchmark #21 in the M5 guidelines).
   (f) Compare the in-sample (training) vs. out-of-sample (testing) fit of the models.
   (g) Do you find similarities in terms of forecast performance across stores or types of items?
3. Consider varying the horizon, h, and obtaining forecasts for $h = 1, ..., 28$. Do you find different rankings of models according to the RMSSE (Root Mean Square Scaled Error) suggested in the M5 guidelines.

4. Now, in this question only, aggregate the data at the store or type of item level.

   (a) Produce forecasts for these aggregates using the previous methods and assess them.

   (b) Compare the forecasts of the aggregates to the aggregates of the forecasts that you had obtained in previous questions, can you find systematic pattern in terms of relative forecasting performance?

5. Now, in this question only, aggregate the data at the weekly frequency,

   (a) produce forecasts for the weekly aggregates using the previous methods and assess them

   (b) Compare the forecasts of the weekly aggregates vs. the weekly aggregates of the daily forecasts, can you find systematic pattern?

6. OPTIONAL: Based on your answers to questions 4 and 5 above, see whether using lags of the aggregates can help forecasting the daily data. When using lags of temporal aggregates, this is related to the HAR model à la Corsi (2009).[1] You may want also to consider an ARFIMA(p, d, q) model where $d \in (0, 1)$.

7. We now only focus on forecasting the disaggregates (original data).

   (a) Consider the probabilistic forecasts i to vi in the M5 guidelines

   (b) Assess them using the Scaled Pinball Loss (p7 of the M5 guidelines).

2 Multivariate Models

We now consider fitting multivariate models and, in each case, comparing them with the univariate benchmarks obtained in Section 1 as well as between themselves. Assessments are made at the $h = 1$ horizon for point & probabilistic forecasts.

1. Consider a VAR model for each store and obtain resulting forecasts for the testing sample

2. Consider a VAR model for each type of products and obtain forecasts.

3. Consider a large VAR model involving all variables.

4. OPTIONAL

[1] Journal of Financial Econometrics, 2009, Vol. 7, No. 2, 174–196; see http://public.econ.duke.edu/~get/browse/courses/672/Lectures/10_AR-HARmodels.pdf for a simple explanation, where the daily data is recorded over a 5-working-day week(so there are about 22 days per month. Here we would need to compute the monthly average over 30 days.

(a) If you wish, you may consider dynamic factor models where you allow the possibility of one factor for each store or each type of item.

(b) If you wish, you may use the whole dataset available on github and consider a large dynamic factor model.

3 Machine/Deep Learning models and extension

1. Now consider machine learning benchmarks, such as those in the set of benchmarks or that are available on the M5 github (e.g. LSTM, using keras in R) and compare their performance in terms of point/probabilistic forecasts.

2. Can combinations of ML/DL techniques with standard techniques help?