## Learning from Multiple Time Series (Exploiting the Hierarchy)

1. The winning method will utilize cross-learning, and global and hybrid models will dominate local models.

In our review of the six competitions, we found that all of the approaches of the top placers across the competitions shared one characteristic: they all used cross-learning. Cross-learning is the usage of data from multiple, often related, time series in model estimation. This addresses some of the key challenges in forecasting, namely that time series are often short and their dynamics change over time. This makes it difficult to estimate complex models, which is why traditionally simple models have been preferred. The logic behind cross-learning is that time series, especially if related, often have similar patterns. This can be leveraged to model effects that one might not otherwise be able to model reliably due to the small sample size. Good examples include modelling of seasonality and holidays with only a few seasons of data, or new promotion strategies and pricing strategies. Different degrees of cross-learning exist depending on how many time series are used in model estimation and it can be viewed as a spectrum. On one end, local models exist, which use only the time series itself and on the other end, global models, where one model is estimated for all time series in forecasting task. In between exists approaches, where one model is estimated for a subset of all time series, e.g. a product family, geographical area etc. The goal is to strike a balance between the amount of data available and the similarity of the time series. In the six competitions, the top placers used either global models or hybrid approaches, where the model had both global and local aspects.

1. Access to hierarchy information will increase the performance gap between local models and models using cross-learning compared to the M4 competition.

In the M4 competition, the top two placers also used cross-learning with global characteristics. This showcased that cross-learning can be effective even for seemingly unrelated time series. In the competitions and in many business forecasting tasks, time series are organized in hierarchy where items at the same level of the hierarchy are related. This information can be used to increase the effectiveness of cross learning by informing the model on which time series are related and expected to have similar patterns. In the competitions where information on hierarchy was available, contestants using cross-learning outperformed local time series models by a larger margin than in the M4 competition. We therefore expect to observe a larger gap between local and global models in the M5 competition than in the M4 competition.

1. Contestants will develop innovative strategies to tackle the challenge of hierarchical forecasting, and we expect new neural network architectures and GBDT strategies to utilize this information optimally.

While the previous Kaggle competitions provided access to hierarchy information, they only required forecasts for one level of the hierarchy. The M5 competition is a true hierarchical forecasting task in that it requires coherent forecasts at all levels of the hierarchy. To the best of our knowledge, this is the first time that a Kaggle competition features hierarchical forecasts. We therefore expect that contestants will develop new ways of adapting their favorite methods, which on Kaggle are often ML methods, to the task of hierarchical forecasting.

## The Rise of Machine Learning

1. GBDT using feature engineering based on e.g., rolling statistics and neural networks will both perform well in the competition and outperform existing time series benchmarks in terms of both accuracy and uncertainty.

Methods not commonly used in time series forecasting dominated the latest four of the competitions reviewed. Interestingly, the two methods were those that also placed in the top two of the M4 competition: neural networks and gradient boosted decision trees. Neural networks have been the topic of much research recently with companies such as Amazon using them successfully at scale, and their performance therefore does not come as a huge surprise. The network designs (known as architectures) used in the competitions varies considerably from models explicitly designed for time series (recurrent or convolutional neural networks), similar in nature to Amazon’s DeepAR, and others closer to a complex non-linear regression model (feedforward neural networks). Their commonality lies in the ability to incorporate exogenous variables in the models. Gradient boosted decision trees is a newcomer to the forecasting scene, and has roots in both ML and statistics literature. They consist of a large number of decision trees grown in sequence, where each tree tries to correct the errors made by the previous trees. This approach generally provides higher accuracy than a single decision tree, but comes at the cost of a complex model. Gradient boosted decision trees are a general regression and classification tool that has shown state-of-the-art results on many prediction tasks, but they have not designed with forecasting in mind. They were used by the 2nd place in the M4 competition to estimate combination weights for local time series models, whereas they have been used to forecast the time series directly on Kaggle. A key to their success for forecasting has been adapting them to time series by providing it with inputs that describe the patterns in the time series in addition to the exogenous variables. Strategies used successfully are similar across the competitions and includes rolling statistics, such as averages, medians and standard deviations, over different time windows at different levels of the hierarchy. It is also helpful to calculate these rolling statistics segmented by day of week, promotions or other exogenous variables known to affect the time series. Due to their consistent strong performance in the Kaggle competitions with similar data characteristics, we expect them to outperform time series benchmarks in the M5 competition as well.

1. 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, which has been the topic of much recent research.
2. Hold-out datasets or time series cross-validation will be used by top placers to avoid overfitting.

## Forecast Combinations Once Again

1. Ensembles of methods will continue to take up the top slots, as is consistent with the findings from all Kaggle and M-competitions. We expect these ensembles to contain both neural networks and GBDT, potentially in combination with other methods.

## External Variables: It Depends

1. Using exogenous variables such as prices, promotions, holidays, and other events will provide improvements to forecast accuracy, in line with previous retail research (Fildes et al., 2019) and the Kaggle competitions.

All of the reviewed competitions provided access to exogenous variables thought to contain predictive power. While these exogenous variables generally proved useful in improving forecast accuracy over that of pure time series methods, this was not always the case. Exogenous variables known for directly affecting the time series within a domain, such as promotions or holidays within retail or reservations within the restaurant domain, proved highly useful. On the other hand, macroeconomic indicators such as consumer price index, oil prices etc. did not prove very useful. This was also the case for weather information, which only proved slightly useful, despite the availability of the actual historical weather data rather than an uncertain forecast. Thus, it seems indirectly related time series or time series that have to be forecasted themselves do not provide much predictive power. We therefore expect that directly influential and known-in-advance information in the retail domain, such as prices, promotions, holidays and events, will help improve forecast performance over that of pure time series methods in the M5 competition.

Current Status

At the time of writing, the M5 competition still has roughly three months to go. While this leaves plenty of time for things to change, the current standing might indicate whether some of our hypotheses are likely to hold true. In terms of both forecast accuracy and predictions intervals, contestants have already managed to outperform all of the benchmark methods on the public leaderboard. Based on the forum discussions on Kaggle, gradient boosted decision trees with time series feature engineering and cross-learning has been used by many contestants to beat the benchmarks for point forecasts. Some contestants have also shared neural networks that outperform the benchmarks, but these are currently not near the top of the leaderboard. Top performers for prediction intervals have so far been quiet on the forums and it is therefore currently unknown how they have managed to beat the benchmarks. The current status thus leads credibility towards our hypotheses of the superiority of global machine learning models over local time series models, although neural networks are yet to show that they are competitive with gradient boosted decision trees. While a sceptic could attribute the superior performance of these methods to overfitting, we do not find it likely considering the results of the earlier similar competitions. On top of this, with three months to go, we expect that large increases in performance are still to come for the top contestants. As for the rest of our hypotheses, it is still too early to speculate on their validity.