Review of Judgemental Selection of Forecasting Models

Fotios Petropoulosa, Nikolaos Kourentzesb, Konstantinos Nikolopoulosc, Enno Siemsen

Harshvardhan
harshvar@vols.utk.edu
University of Tennessee

September 19, 2022

In this paper, Petropoulos et al. (2018) present the results of a behavioural experiment to showcase the pros and cons of using human judgement in choosing the best forecast. In general, forecasting algorithms and packages provide several options to the practitioner. The most common examples are time-series models such as forms of exponential smoothing models (ETS), autoregressive integrated moving averages (ARIMA) and modern machine learning models such as decision trees and neural nets.

With the proliferation of models, developing a framework for choosing the best model is essential. A typical forecasting software package such as SAS picks the best package based on statistical algorithms. These algorithms determine the best performing model on either the in-sample or out-sample data. In the former case, the information provided by the forecast is judged using information criteria such as AIC and BIC. In the latter case, the forecast is ruled by the model's accuracy in the holdout set. This is commonly known as cross-validation.

One of the fundamental results of this paper is that averaging results across forecasts perform better than any singular model. Their behavioural experiments found that 50-50 weight between managerial and statistical forecasts, commonly referred to as **?**, routinely outperforms either forecast. This is significant for our work with HP as we could create an ensemble of statistical, consensus and ML models. Even with equal weights, we could develop better forecasts than any model could do on its own.

1 Judgement in Forecasting

Practitioners use judgement in forecasting at several steps. The authors enumerate five crucial steps in the process that require clear judgement.

1. Defining candidate set of model: which models to consider in the overall planning,

2. Selection of a model: how to choose the model to use,

3. Parametrization of model: tuning parameters of the chosen model to enhance its performance,

4. Production of forecast: creating complete forecasts to be used by planners. This might involve aggregation or disaggregation as most forecasts in real-world is hierarchal (Abolghasemi et al., 2019),
5. Forecast revisions/adjustments: forecasts frequently require updates from managers and experts at the company. Such information is called "soft data".

Experts are often asked to adjust (or correct) the estimates provided by the statistical methods to take additional information into account. Past studies such as Bunn and Wright (1991) called choosing the model (step 2) as "judgemental model selection". In this work, the authors identify a process that most likely will choose models that lead to improved forecasting performance.

2 Common Methods of Forecasting

Traditional Methods Most business forecasting methodologies are based on simple, univariate models. The most popular model is called exponential smoothing, often abbreviated as ETS (for ExponenTial Smoothing, or Error, Trend, and Seasonality). The error term could be additive (A) or multiplicative (M). The trend and seasonality could be missing or none (N), additive (A), or multiplicative (M). The trend could be linear ($d = 1$), damped ($d < 1$) or accelerating ($d > 1$).

Exponential smoothing models are one of the most successful forecasting methods. They produce forecasts using a weighted average of past observations, with the weights decaying exponentially as the observations get older. That is, more recent observations have higher weights.

Non-traditional Methods There are alternative methods to forecasting, such as neural networks and machine learning forecasts. These models are better at data fitting and pattern recognition than traditional models and thus often have higher accuracy metrics. However, machine learning forecasts are relatively nascent and have not proved their mettle in the field. Such complex methods are often considered black boxes though measures such as SHAP scores can help interpret (Lundberg and Lee, 2017).

3 Model Selection Process

The model selection process (item 2 in the list above) could use either only algorithms or algorithms in conjunction with humans or only humans.

3.1 Algorithmic Model Selection

Algorithmic model selection methods use predefined statistical criteria based on which the best model is chosen. Most methods fit the data and create the forecasts, then calculate the forecast error or the information captured in the forecast. Commonly used methods are

Akaike's Information Criteria (AIC) and Bayesian Information Criteria (BIC). AIC after correction for small sample sizes (AICc) is often recommended as the default option.

Information criteria are based on an optimised likelihood function penalised by model complexity. A model with optimal likelihood optimises the one-step-ahead forecasts, assuming that the resulting model parameters would also be optimal for more extended horizon error distributions.

This assumption is not certainly true. It is possible and even expected that the error distribution is often time-dependent. Mathematically, the models assume $\epsilon^2 = \mathscr{N}(0, \sigma^2)$ while in truth $\epsilon^2 = \mathscr{N}(0, \sigma_t^2)$ is observed. When practitioners do not recognise this, they risk choosing a biased model favouring one-step-ahead performance at the expense of longer time horizons.

In the original implementation of the model described above at HP, they used the model with the most minor mean squared error as the final machine learning model. How do the forecasters choose their models? That's not known. We notice that the Consensus forecast doesn't change every month. It could be so, as Cara says, because they forecast in quarters, scale down to monthly buckets, and do not update their monthly forecasts, while Statistical and ML forecasts are updated.

3.2 Validation Set

Another alternative is using a validation set to choose the best model. The available data set is divided into a training and a validation set. The model is created using the first set and evaluated on the second set. The model with the highest performance accuracy on the second set is considered the best model.

The decision maker can choose the appropriate accuracy measure. Selecting the proper accuracy measure as the actual cost function used to train the model is preferable. The forecasts for validation may be produced only once (fixed-origin validation) or multiple times (rolling-origin), which is the cross-validation equivalent for time-series data. Evaluating forecasts over rolling origins has advantages, notably as they are robust to one-off peculiarities in the data, which affect a single validation window (Tashman, 2000).

Model selection based on a validation set has two significant advantages over information criterion-based selection. First, the performance of multi-step ahead forecasts can be used to inform model selection itself. Second, the validation set can be used to evaluate any number of models, including an ensemble of models. Its disadvantage is that it requires a holdout set which may not be feasible for new products.

3.3 Agency Model Selection

In the third case, managers choose the model to use based on some rules. These rules involve measuring various time series characteristics such as trend, seasonality, skewness, intermittence, variability, etc., as well as decision variables such as forecast horizon.

It is also possible that the rules are not formally defined but only operates on the managers' intuitions. At HP, the ML model only produced forecasts for the next six months, while Stat and Consensus forecasts were created for 18 months (though it is trivial to expand).

Managers often believe that they understand the business contexts better and would occasionally override the algorithmic forecasting rules. For example, Petropoulos et al. (2018) write, "even if the result of an algorithm suggests that the data lacks any apparent seasonality (and as such a seasonal model would not be appropriate), managers may still manually select a seasonal model be- cause they believe that this better represents the reality of their business."

At HP, it wasn't uncommon for forecasts to reflect company targets instead of actual algorithm-based predictions. For example, the company would like a product to succeed and inflate its forecast to reach the goal. Also, for the products at the end of their life cycle, even though the algorithm proposes a forecast, business sense would manually decrease the numbers to make it make sense.

4 Combination and Aggregation

Armstrong (2001) found that forecast combination often significantly improves forecast accuracy. Blattberg and Hoch (1990) used a simple 50-50 combination and saw significant gains in the combination methods than any other algorithm individually. Their results have repeatedly been replicated by many.

Franses and Legerstee (2011) found that a simple combination of forecasts outperformed both statistical and judgmentally adjusted forecasts. Petropoulos et al. (2016) demonstrated that a 50-50 mixture of estimates in the period after a manager's adjustments have resulted in significant losses can indeed increase accuracy by 14%. Wang and Petropoulos (2016) found that a combination is as good, if not better than selecting between a statistical or an expert forecast. Trapero et al. (2013) demonstrated further gains with more complex combinations.

5 Behavioural Experiment

The authors performed a behavioural experiment where they provided the option of choosing the best forecast to 693 participants (276 UG students, 211 PG students, 44 researchers, 90 practitioners and 72 others). The participants were divided into two groups. The first group was presented with four options and had to choose the best one by visually examining them. This approach was called "judgemental model selection". The second group was presented with an option where the user partially built the model by identifying trends and seasonality in the model. Based on the choice, one of the four models was automatically selected. This was called the "model-build" approach.

The four model options presented were four forms of exponential smoothing models (ETS): (1) Simple exponential smoothing (SES), (2) SES with additive seasonality, (3) Damped exponential smoothing (DES), and (4) DES with additive seasonality.

A user's choice of model was used to forecast accuracy over a twelve-month horizon. The data was sourced from the M3-Competition dataset (Makridakis and Hibon, 2000). 32 time series were handpicked with a monthly frequency such that for 16 of them, AIC-based model selection would identify the best model, and for another 16, AIC-based model selection would not have identified the best model.

5.1 Forecast Measures of Accuracy

The authors used the following measures to measure forecast accuracy: MPE, MAPE and MASE.

$$MPE = \frac{100}{H} \sum_{i=1}^{H} \frac{y_{n+i} - \hat{y}_{n+i}}{y_{n+i}} \tag{1}$$

$$MAPE = \frac{100}{H} \sum_{i=1}^{H} \frac{|y_{n+i} - \hat{y}_{n+i}|}{|y_{n+i}|} \tag{2}$$

$$MASE = \frac{(n-1)\sum_{i=1}^{H} |y_{n+i} - \hat{y}_{n+i}|}{H \sum_{j=2}^{n} |y_j - y_{j-1}|} \tag{3}$$

5.2 Results

The authors found the following conclusions.

1. Participants performed better in the "model-build" approach than in the "model select" approach. Thus, thinking in components builds better models than thinking absolutely.

2. Humans are superior in avoiding the worst model compared to using algorithms (i.e. choosing models based on AIC) to pick the best model.

3. 50% statistics + 50% judgement performed better than statistical models alone. This model was less biased than statistical selection in 86% and produced lower values for MAPE and MASE in 99% and 90% of the cases. This encourages the use of ensemble models at HP.

4. Wisdom of crowds: Aggregating five people's opinions produced a better model than any one model individually. This is expected – regression to the mean.

5. A weighted combination of models based on AIC improves the performance beyond the statistical model. However, this was always outperformed by 50-50 combinations of statistical and judgemental forecasts and by the wisdom of the crowds.

The authors also propose a 3 × 3 matrix of Forecastability and Importance (Low, Medium and High). Those with high importance and low forecastability could use judgemental forecasting most.

References

Abolghasemi, M., Hyndman, R. J., Tarr, G., and Bergmeir, C. (2019). Machine learning applications in time series hierarchical forecasting.

Armstrong, J. S. (2001). Combining forecasts. In *Principles of forecasting*, pages 417–439. Springer.

Blattberg, R. C. and Hoch, S. J. (1990). Database models and managerial intuition: 50% model + 50% manager. *Management Science*, 36(8):887–899.

Bunn, D. and Wright, G. (1991). Interaction of judgemental and statistical forecasting methods: issues & analysis. *Management science*, 37(5):501–518.

Franses, P. H. and Legerstee, R. (2011). Combining sku-level sales forecasts from models and experts. *Expert Systems with Applications*, 38(3):2365–2370.

Lundberg, S. and Lee, S.-I. (2017). A unified approach to interpreting model predictions.

Makridakis, S. and Hibon, M. (2000). The m3-competition: results, conclusions and implications. *International journal of forecasting*, 16(4):451–476.

Petropoulos, F., Fildes, R., and Goodwin, P. (2016). Do big losses in judgmental adjustments to statistical forecasts affect experts behaviour? *European Journal of Operational Research*, 249(3):842–852.

Petropoulos, F., Kourentzes, N., Nikolopoulos, K., and Siemsen, E. (2018). Judgmental selection of forecasting models. *Journal of Operations Management*, 60:34–46.

Tashman, L. J. (2000). Out-of-sample tests of forecasting accuracy: an analysis and review. *International journal of forecasting*, 16(4):437–450.

Trapero, J. R., Pedregal, D. J., Fildes, R., and Kourentzes, N. (2013). Analysis of judgmental adjustments in the presence of promotions. *International Journal of Forecasting*, 29(2):234–243.

Wang, X. and Petropoulos, F. (2016). To select or to combine? the inventory performance of model and expert forecasts. *International Journal of Production Research*, 54(17):5271–5282.