Judgemental Selection of Forecasting Models

My Notes

Harshvardhan
harshvar@vols.utk.edu
University of Tennessee

August 30, 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 Blattberg and Hoch (2010), 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.

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 aforesaid model at HP, they used the model with the least error as the final machine learning model. The actual criterion used by the forecasters is not known.

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.

– more details here

There is an ongoing discussion on which metrics to choose for the best model. For details, see Section 4.3 of "My Learnings from Internship at HP Inc."

Agency Model Selection The managers choose the appropriate model in the second case.

References

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

Blattberg, R. C. and Hoch, S. J. (2010). Database models and managerial intuition: 50% model+ 50% manager. In *Perspectives On Promotion And Database Marketing: The Collected Works of Robert C Blattberg*, pages 215–227. World Scientific.

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

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

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