Warning: Do Not Just Average Predictions!

By Ville Satopaa , INSEAD Assistant Professor of Technology and Operations Management

A novel framework for understanding and aggregating multiple predictions from diverse sources.

At a 1906 livestock show in Plymouth, England, nearly 800 people participated in a **contest** to guess the weight of a slaughtered ox. The average of these estimates was 1,197 pounds. This is remarkable because the true weight of the ox turned out to be 1,198 pounds. The average was only one pound away from the truth! How could it be so accurate? Perhaps by chance?

Today it is generally agreed that combining multiple predictions or estimates leads to more accurate forecasting than merely relying on a single prediction. Academic research, however, does not agree on the best way to do this. For instance, imagine asking 1,000 people to predict the demand of a new product that you are about to launch. Suppose these predictions vary from as low as 100 units to more than 10,000 units. How can such different opinions be summarised into a single consensus that represents the best prediction out there? Would averaging this type of prediction work as well as

it did in the ox example? To answer this, we must revisit some foundational assumptions of statistics. More specifically, we must understand why these predictions differ from each other and from the true demand. Only then we have some hope of finding a principled summary that is useful to managers.

Co-written with Robin Pemantle and Lyle Ungar, our article, **"Modeling Probability Forecasts via Information Diversity"** in the *Journal of the American Statistical Association*, does exactly this. The paper returns to the fundamentals, develops a general modelling framework called the partial information framework, and then answers the question: Should one use an average to summarise different predictions? Surprisingly, the answer is almost always "no".

Why average?

Historically, statistics have explained differences among observations and the truth with **measurement error**. The classic example of this considers a person repeatedly measuring the height of some object, say, a book with a ruler. Every measurement is likely to be slightly different because the person is unlikely to carry out the measuring exactly the same way twice. Sometimes the measurement is higher and sometimes it is lower than the true height of the book. If the person is equally likely to over- and under-shoot, the true height will be close to the middle of the measurements. To capture this central value, a natural and theoretically sound approach is to average the measurements. Perhaps due to its simplicity and familiarity, however, averaging is often applied far outside its proper context and even when measurement error is not the dominant driver of the differences.

Partial information framework

In the book example, the observations came from a single instrument, namely the person with the ruler. In the demand forecasting example, however, the observations came from 1,000 different people or "instruments''. Now the observations do not differ only because of errors in measurement but also because the instruments themselves are different. Relying on the classical measurement error ideas and hence simply averaging the predictions would ignore the inter-person variation. This leads to poor forecasting accuracy because, as our paper explains, the inter-person variation is often more important than measurement error and hence should be the key driver in choosing the appropriate way to summarise the predictions.

Our *partial information framework* addresses this with both aggregation and analysis of predictions from multiple sources. The underlying approach incorporates and understands the information that each person is using in the prediction, which isn't possible in averaging. The framework explains and mathematically formalises the differences in the individual predictions with information diversity. The framework makes no assumptions about the sources of the forecasters' information. For instance, the information could stem from books, movies, personal recollections, interviews or any other source.

Under this framework it is then possible to combine the predictions into a consensus that reflects the information of the entire group. For instance, consider two people making predictions. Person 1 knows facts A and B whereas Person 2 knows facts B and C. The consensus is then a forecast that efficiently uses all facts A, B and C. The idea is quite straightforward when there are only two forecasters – there is some amount of information that Person 1 only knows and some amount that Person 2 only knows and there's some overlap. But if there are 100 people, the network of overlapping information becomes quite complicated. For instance, the first two people may both know something that the third person does not and so on. Fortunately, the *partial information framework* can incorporate any such overlapping sets and find a consensus that reflects the total information among any number of forecasters.

When averaging is not enough

The *partial information framework* is not only useful for combining predictions. It can also be used for understanding many commonly used ad-hoc approaches. To this end, our paper analyses an empirical technique known as *extremizing* that pushes the average forecast to one end or the other, away from the middle. For example, suppose ten people estimated the chances of Marine Le Pen winning the last French presidential election. If the average of these forecasts is 0.2 (or 0.75), extremizing would push that probability closer to 0.0 (or 1.0, respectively). This technique has been observed to improve the average forecast in many different applications, including political foresight, weather forecasting and others. It has been less clear, however, when one should extremize heavily and when not at all. Our analysis now shows that if all experts have exactly the same information (and can then be considered as the same instrument), do not extremize at all; just take the average. However, if they all rely on completely different

information, that is, everything that one person knows, another does not, then extremize heavily.

Imagine two people predicting a coin toss. Person 1 magically knows the result every time but Person 2 has no idea – after all, it's a coin toss. Person 2 reasonably says the probability of heads is 0.5 for every coin toss and Person 1 says 0.0 (definitely heads) or 1.0 (definitely tails) each time because that person knows the result. The average of these predictions is either 0.25 or 0.75, depending on what Person 1 predicts. Clearly this is not a good estimate. Extremizing would help because it would push 0.75 closer to 1.0 or 0.25 closer to 0.0. In this instance, the information sets are completely disjoint. Therefore the average forecast should be extremized heavily.

Next consider our first example about the slaughtered ox. Here the average estimate is remarkably accurate and hence requires no extremization. This makes sense because at the fair, each participant had the same level of access to the ox. In other words, all relevant information was equally available to everyone. Therefore the participants were using the same information, and no extremization was needed.

These two examples illustrate the ends of the spectrum. Most real-world forecasting scenarios fall somewhere in between. In such cases, some extremization is helpful, and simply averaging the forecasts is not enough.

Previously, extremizing has been used in an ad-hoc manner. The amount of extremizing has been learned based on past data (past forecasts for which the true outcome is already known). But now we have a mathematical model, namely the *partial information framework*. Our framework motivates and leads to techniques that "automatically'' extremize the average forecast just the right amount based on the forecasters' information overlap. Overall, this is more principled and does not require any past data, making it applicable even when ad-hoc empirical techniques, such as extremization, are not.

More than humans

Our *partial information framework* is not constrained to predictions made by people. Predictions could come from machines, people or both. In particular, a joint analysis of human and machine predictions could help us better understand when human forecasters can outperform machines (or vice versa).

Moving away from the measurement error model for forecasting has been on the cards since the 1960s. Our novel approach finally offers help to forecasters and also opens up many avenues for future research. Overall, the *partial information framework* is a very general modelling framework that can be used to analyse real-world forecasting data in many different ways. Similarly to classical statistics, one must choose a distribution (normal/Gaussian distribution, t-distribution or others) before applying the techniques. Each choice leads to a different partial information model; which model works best depends on the application at hand. Therefore companies ought to collect forecasting data and perform large-scale model selection to find the best partial information model for their application. This model can then be used to construct the optimal way to combine the predictions in the future. Such a principled approach does not only describe the uncertainty in the consensus prediction but is also likely to provide dramatic improvements in the overall accuracy.

Ville Satopää *is an Assistant Professor of Technology and Operations Management at INSEAD.*

Find article at

https://knowledge.insead.edu/operations/warning-do-not-just-average-predictions

About the author(s)

Ville Satopaa is an Assistant Professor of Technology and Operations Management at INSEAD.