On the adverse consequences of perceived model complexity

Models play an ever more important role in the financial industry. Indeed, models at banks are being used for a large variety of different purposes, ranging from pricing loans and hedging derivatives to transaction monitoring in the context of detecting financial crime. One of the key challenges in building these models is finding the sweet spot between simple models and complex models. In that light, parsimonious model selection methods such as Occam's razor have become increasingly popular tools for model developers in financial institutions. However, properly applying these principles of parsimony is often made difficult due to the discrepancy between perceived complexity and *actual* complexity.

In many financial institutions, key decision-makers in upper management (oftentimes without a quantitative background), prefer models that they perceive to be simpler over models which they perceive to be more complex. At first glance this inclination seems to be in line with commonly used parsimonious model selection principles such as Occam's razor or minimum description length. In many cases, however, the model that is perceived to be more complex in reality contains an amount of parameters that is equal to (or in some cases: *lower than*) the model that is perceived to be simpler. Next to that, the (supposedly) more complex models are often easier to interpret and implement, capture more sophisticated dynamics and require fewer assumptions than their supposedly 'simpler' counterparts. Hence, the discrepancy between perceived and actual complexity often leads to a sub-optimal model choice. To demonstrate this fact, let us first consider a simple example of this phenomenon in the modelling of the loss-given-loss (LGL) and then a more complicated example of this phenomenon in transaction monitoring in the context of detecting financial crime.

The key risk driver included within LGL modelling is the so-called loan-to-value (LTV). Here, the LTV measures the ratio between the principal amount, and the value of the underlying collateral, i.e. LTV = principal amount / value of the collateral. The higher the LTV, the higher the expected losses (which is also clearly reflected in the pricing of retail mortgages). When the size of the loan is large compared to the value of the underlying collateral, the recoveries resulting from the sale of the collateral may be insufficient to fully cover the outstanding.

Having established that LTV is a key risk driver in the context of LGL modelling, it would only be natural to consider an LGL-model of the following kind:

$$LGL = a\ LTV + b$$

where $a \in R^{+}$ and $b \in R$. At first glance, the choice for this model formulation seems well justified. For one, this formulation captures the required dynamic: the higher the LTV the higher the expected loss on the loan (considering the fact that $a \in R^{+}$). More importantly, by using such an intuitive linear relationship between LGL and LTV it also seems that we act in line with the principle of parsimony. The benefits of such a simple formulation are manifold. For instance, such a model structure is relatively straightforward to explain to non-quant stakeholders. On the quantitative side, note that simpler models oftentimes reduce the risk of overfit, and also allow us to rely on ordinary least squares estimation. Hence, it is no surprise that the above formulation is being used by many financial institutions at the time of writing this article. However, looks can be deceiving.

E.D. Tauran MSc is consultant at RiskQuest.

G.C.P. van Miert PhD is senior consultant at RiskQuest.

S.P.J. de Man PhD is Partner at RiskQuest.

Upon further inspection of this model formulation, one quickly learns that there are some dynamics that this model formulation does not allow for. For instance, the above model can easily give rise to an LGL which is much larger than 100% or lower than 0%, which (disregarding additional drawings on the loan, etc.) is nonsensical from a business point of view. Hence, in practice one is forced to replace the linear model by the slightly more convoluted formulation:

$$LGL = \max\{0, \min\{a\ LTV + b, 1\}\}.$$

Note also that in this formulation, apart from their statistical meaning there is no business interpretation for both the a and b parameter. Next to that, experience learns that oftentimes the relation between LGL and LTV is not completely linear and hence this formulation often does not correspond well with the actual structure of the data.

Considering the fact that upon further inspection this simple model formulation seems to have several drawbacks, let us take one step back and note again that the LTV is just the loan divided by the value of the collateral and that the LGL is just the loss divided by the loan, that is,

$$LTV = \frac{loan}{value} \text{ and } LGL = \frac{loss}{loan}.$$

Note also that in the case that the collateral is sold, the loss on the loan hence becomes

$$loss = loan - value * SR$$

where SR can be interpreted as the sales ratio of the collateral (ratio of realized sale price of collateral to intrinsic value of collateral). Upon dividing on the left- and right-hand sides by the loan amount we find:

$$LGL = 1 - \frac{value}{loan} * SR = 1 - \frac{SR}{LTV}$$

To also account for recoveries resulting from other sources, such as wage garnishment one will also typically include an additional constant c:

$$LGL = 1 - \frac{SR}{LTV} - c$$

What we have now obtained is a model formulation for the LGL that at first glance might seem more complex – purely because it is non-linear – but in practice turns out to be much simpler than its linear counterpart. Note first that in terms of the amount of parameters that need to be estimated, the perceived simple formulation requires the estimation of two, namely the a and the b parameter. This is equal to the amount of parameters that need to be estimated in the non-linear case, which requires estimation of the SR parameter and constant c. Secondly, note that in the first formulation, both the a and b parameter have no natural interpretation whereas in the elegant formulation the SR parameter can be interpreted as the sales ratio of the collateral of the loan. Hence in this case the elegant formulation of the model has the added benefit of being explainable to stakeholders due to having interpretable parameters. Next to that, we find that in contrast with the simple formulation, the elegant formulation allows for correct limit behavior as when the value of the collateral of the loan goes to zero, we find that the term in the LGL accounting for the collateral goes to zero.

Given all of these advantages, it is very surprising that the simple formulation is still used so often by financial institutions. What seems particularly interesting about this example is that the simple fact that throughout the financial sector this familiarity with LTV can have such drastic consequences. If instead, the value-to-loan (VTL) would have been the acronym of choice we expect that it would have been much easier to convince stakeholders of the appropriateness of the second formulation over the first as then the second formulation would have been linear.[1] Hence, this example clearly teaches that by itself a non-linear relationship is not more complex, it is only perceived as such, owing to the popularity of LTV over VTL.[2]

Another domain in which perceived complexity results in sub-optimal model choices appears in the field of transaction monitoring in the context of detecting financial crime. Many financial institutions opt for the business-rule approach, which essentially boils down to formulating and maintaining a collection of if-statements that trigger alerts whenever certain pre-specified conditions are met. At first glance, this approach seems simple and explainable. Every individual business rule is simple, uniquely defined by a set of conditions and serves to mitigate a specific risk. This makes for easy communication to stakeholders since formulating a sufficient amount of business rules will then supposedly cover the entire spectrum of financial crime.

In practice however, formulating and maintaining these business rules quickly becomes a cumbersome endeavor that does not seem to be very scalable since – unfortunately – the transaction behavior of financial criminals changes over time. Additionally, as the number of business rules increases it becomes increasingly hard to maintain an overview of the aggregate performance of this now substantial collection of business rules.

Oftentimes, model developers at financial institutions suggest using machine learning methods as an alternative method for detecting unusual transactions. In contrast to the (perceived) simple business rules, machine learning methods are often perceived as complex by managers because of their mathematical sophistication, black-box nature and corresponding lack of explicability. In terms of effectiveness, scalability and maintenance requirements however they are far superior to business rules when it comes to detecting unusual transactions in the battle against financial crime.

All considered, matters that appear to be simple at first glance unfortunately turn out to be more convoluted than initially expected, and vice versa. In that light, it is of paramount importance that key decision-makers at financial institutions trust model developers in making the right choices with respect to model selection. Note that this does in no way absolve model developers of their responsibility to clearly communicate their considerations with respect to model selection. Clear communication is an essential part of the model development process, and fortunately so, the best cure against the adverse consequences of perceived complexity. ■

1 – In that sense, we should consider ourselves fortunate that it is not market practice to use the hyperbolic tangent of the LTV.

2 – For those who remain unconvinced, note that the 1/LTV formulation is also supported by the capital requirements for credit risk within the Solvency II framework.