Opinion pieces

On data and models: Is more always better?

Received (in revised form): 2nd October, 2023

Thomas C. Wilson
Professor, Ludwig Maximilians University; CEO, Allianz Ayudhya Pcl., Thailand

Thomas C. Wilson is an honorary professor of applied risk in insurance at Ludwig Maximilians University, and Chief Executive Officer of Allianz Ayudhya Pcl., Thailand. Prior to his appointment in 2020, Tom was Chief Risk Officer (CRO) of Allianz Group from 2008, CRO of ING Insurance from 2005, Global Head, Oliver Wyman Finance & Risk Practice from 2002, Chief Financial Officer of Swiss Re New Markets from 1998 and Global Head of McKinsey Risk Management Practice from 1990. Tom is widely published and author of 'Value and Capital Management: A Handbook for the Finance and Risk Functions of Financial Institutions'. Tom earned his BSc in business from U.C. Berkeley and a PhD in economics from Stanford.

Allianz Ayudhya Pcl., Ploenchit Tower, 898 Ploenchit Road, Lumpini, Pathumwan, Bangkok, 10330, Thailand
E-mail: thomas.w@allianz.co.th

Abstract With regards to data and model sophistication, the new mantra for financial services and FinTech seems to be 'the more the better', supported by attractive business cases in risk underwriting, fraud detection, customer lifetime value management, conditional investment risk/return optimisation and improving customer journeys, to name a few. However, more data and more sophisticated models are not always a universal panacea and may lead to bad business outcomes if not managed appropriately in the context of the desired business outcome. This paper summarises the evolving business cases for increasing data and models in the risk management domain and their associated risks. Making the associated risks transparent naturally leads to the conclusion that a timeless risk mitigation approach — common sense — is critically necessary to complement the more structured model risk management (MRM) framework that is evolving.

Keywords: *data, models, risk management, insurance, analytics, artificial intelligence*

INTRODUCTION

This paper addresses the increasing value of data and analytics to risk professionals. Increasing data and model complexity have been driven by five factors:

- increased data availability;
- more advanced modelling approaches;
- cheaper and more available memory and computational 'horsepower';
- regulatory acceptance (and promotion) of risk-based models; and
- financial agents' self-interest in improving personal and corporate outcomes.

The first three of these trends are at the foundation of the terms 'big data' and 'big data analytics'. Big data can be characterised as data which is beyond the scope of relational databases to manage and analyse due to 3-Vs[1], eg a broader *variety* of data types including both structured and non-structured data, delivered in a higher *volume* and at a higher *velocity*, including real time data. Big data analytics[2] can be characterised as the process of deriving trends, patterns and correlations from big data with the goal of understanding what happened in the past and why (descriptive and diagnostic analysis), what could happen next (predictive analysis) and what might happen in the future if certain actions are taken (prescriptive analysis).

While it is intuitive that risk management benefits from increased data and improved analytics, less clear are the potential drawbacks and risks. This contribution tempers the exciting business use cases which leverage data and complex models by elaborating some commonsense principles and caveats borne out of practical experience.

THE BENEFITS

Speaking apocryphally, many believe 'more is better' regarding data and model complexity, equating data as the 'new oil' fuelling advanced artificial intelligence (AI) models, bringing about a bright, post-industrial era. Financial service professionals are not an exception: incumbents (and emerging FinTechs) are making large investments in data and analytics with two broad objectives:

- Improving the understanding of (conditional) probability distributions to support business decisions and assess risk. Examples include underwriting propensity models (eg for loans or insurance policies, for prepayment or lapsation, for up-/cross-sell, etc) as well as conditional value at risk and expected investment risk/return models.
- Improving customer journeys, making them faster, more efficient and more customer friendly. Examples include AI-enabled image recognition systems for claims handling, interactive chatbots and AI-supported self-servicing options, etc.

As illustrated in Table 1, these objectives are met in practice by a wide variety of business use cases leveraging data and analytics.

Table 1: Business use cases leveraging data and analytics

Goal	Business use cases
Improve understanding of (conditional) probability distributions	**Customer lifetime value management** • Improved lead generation and conversion through customer segmentation: What customers are receptive to contact and likely to purchase which product or service? • Improved underwriting decisions: Which loan or insurance policy to approve to balance type 1 errors (approve when you should have rejected) and type 2 errors (reject when you should have approved)? • Improved pricing: How much margin is supportable above the products' technical price given the segments' price elasticity and customer lifetime value? • Improved retention: Which segment is likely to bring the most long-term value through higher utilisation, cross/up-sell, etc? • Improved fraud detection: Which default or claim is likely to be associated with fraud? Using discriminant analysis, structural models (eg Merton credit model), machine learning, AI, etc. **Reserving, stress testing and capital adequacy from a regulatory and business perspective:** • What are the (conditional) mean, median and extreme quantiles of a profit/loss distribution? • What (combined) events are most likely to cause losses? **Investment and capital allocation decisions:** • What are the conditional expected returns of different asset classes or margins of different products? • What is the joint distribution of returns from a portfolio of different asset classes or products? Using multivariate distributional assumptions, structural econometric models, agnostic time series models (ARIMA, GARCH), etc.
Make processes more efficient and customer-friendly	**Leverage structured and non-structured data and models to improve customer journeys**, asking less of the customer but providing faster, more accurate decisions, eg: • Improving the new business submission journey, eg video non-face-to-face know your customer; data completion to support underwriting from eg data bureaus, social media, etc; drone video or IoT (Internet of Things) information for the underwriting of property and agricultural loans and insurance, etc. • Improving insurance claim journeys, eg leveraging claims bureau data, damage photos or drone videos for claims estimates and submission, etc. • Leveraging generative AI and chatbots to improve customer (self-)service options. • Adjusting acceptance, claim and fraud adjudication rules based on models, avoiding more costly and time intensive human intervention. Using AI to interpret and integrate non-structured data into the decision framework and interact with the customer in a cost-efficient manner.

PRINCIPLES

As seen in Table 1, valuable business use cases leveraging data and models are being developed and extended continuously. Recognising the value of data and models leads to a few general principles which govern their use in the risk management domain.

Principle 1: More data is generally better

This is because an increased training sample size:

- increases the out-of-sample accuracy of models and mitigates over-fitting due to too few observations;
- leverages less visible relationships by extending the dataset to include additional information of potential explanatory power;
- reduces potential model bias by extending the dataset to include more segments, allowing for a more accurate view of segment-specific behaviour.

Principle 2: Continually evaluating new models and scientific approaches is generally good

As Table 1 illustrates, not only are new risk management use cases emerging, but current use cases can also be sharpened with the application of increased data, advanced models and computational power.

Principle 3: Data acquisition and modelling efforts should be increased until the marginal benefit equals the marginal cost

- The marginal cost of data acquisition includes the cost of acquisition (eg the cost of scraping data from the Internet, the cost of experiments to elicit consumer responses, the direct cost of second party data providers and the connectivity required, etc), the cost of storage (eg cataloguing and documenting data, memory and archiving costs, etc) and the cost of making big data available (eg analysis and visualisation software, hardware, etc).
- The marginal cost of improved modelling includes the (opportunity) cost of specialist personnel such as AI data scientists with domain expertise as well as the increased computational and storage infrastructure and maintenance cost required by more advanced models.
- The marginal benefit of data acquisition and model improvements will depend on the specific business application (eg in terms of generating revenue, reducing risk, balancing expected returns against downside scenarios, making better decisions, etc) and the efficacy of the data or model in improving the desired outcome.

Very often the marginal benefits are not known *a priori*, implying that a sequential acquisition-model-evaluation cycle is optimal.

THE CAVEATS

There are, however, a few common sense caveats to these principles which, while not sufficient to stifle innovation, nonetheless need to be kept in mind.

Caveat 1: Complex models may answer 'the' question but still not be 'fit for purpose' within a business or public policy context

Using more data and more complex models may produce a model which is not 'fit for purpose' in the context of the overall business objectives or processes. For example, in a hypothetical world, all data is available all the time; unfortunately, this is not the case in the real world. The problem is that complex models may not be robust to missing input data, requiring all inputs in order to yield improved results; however, their efficacy (and accuracy) may deteriorate dramatically relative to more simple models if some input data is missing or corrupted. As an example, consider that men and women have different threshold levels of 'good' high-density lipoprotein (HDL) cholesterol, with a woman's >50 mg/dL and a man's >40 mg/dL. The wrong conclusions may be reached if a health insurance underwriting model is trained using both gender and HDL levels, but gender is missing in the input data for evaluating a specific individual.

Additionally, it is challenging to integrate complex models into business processes if they do not provide results which are intuitive or interpretable to humans. For example, the rationale behind an underwriting, claim or fraud detection decision may need to be disclosed to customers or intermediaries as part of the customer journey or interpretable by management to prevent unconscious bias. Complex AI models which provide an answer without an interpretable rationale may not be suitable for these customer journeys or business processes. As an example, I recently decommissioned a machine learning AI fraud model because I did not want to rely on a machine to make an accusation of fraud without human review. Instead, we switched to a statistical model which gives concrete direction for the claims adjuster to look into.

Furthermore, complex, risk sensitive models may not be robust to extreme events. For example, banks and insurers may be led to assume significant notional exposure based on risk-based models if the modelled risk is insignificant, leaving an overall exposure that can turn catastrophic if a major disruptive event occurs; practical historical examples include market dislocations for 'zero deemed risk' government debt or third party guaranteed derivative positions and 'low probability' accumulation scenarios for collateralized debt obligation (CDO) super senior tranches prior to the 2008 financial crisis, accumulation scenarios for cyber insurance or excess of loss reinsurance covers, to name a few. As another example, models trained on long time series data missed obvious trends in monetary policy.

Finally, more data and advanced models may be in conflict with public policy objectives. For example, in general, underwriting should explicitly exclude proscribed attributes such as gender, religion, race, etc; however, the use of additional data from social media, purchasing patterns defeats, etc, can implicitly defeat this ban. In addition, using more data can preclude some segments from accessing financial services; for example, genetic testing may uncover a genetic predisposition to a chronic disease which might lead rational insurers to a denial or restriction of health insurance cover.

Many of these issues can be mitigated by common sense: first, through the use of simple models to complement more advanced modelling approaches and, secondly, by the explicit recognition that the value of information and models is dependent on the business context.

Caveat 2: The past is not always a good indicator of the future, especially when policy regimes or consumer behaviour change

Models which assume that the past is a good predictor of the future are likely to lead to uncomfortable surprises when the future develops in unexpected ways. In other words, models are trained on finite, past data which might not represent the future, especially during periods of significant discontinuity. The following are two likely sources which may cause dramatic change.

The first is unexpected government policy changes. For example, interest rate models trained during the 1990s were caught flat-footed by Central Bank expansionary monetary policy following the 2008 financial crisis with the resulting lower rates having a dramatic impact on bank net interest income and the asset/liability risk of insurers focusing on retirement savings and investment policies. Similarly, interest rate models trained during 2010–2020 were also caught flat-footed by tightening monetary policy to combat inflation, arguably leading to the recent regional bank crisis in the US and pension squeeze in the UK. Large impacts can also occur from changes in fiscal or tax policy, 'green' subsidies or preferential treatment for alternative asset classes, national security considerations impacting investment decisions, pandemic policies, etc.

Looking at even longer time periods is sometimes not a solution: monetary policy pre-Bretton Woods had a different effect on foreign exchange and interest rate developments across nations than post-Bretton Woods and so concatenating the two datasets may yield worse results. As another example, as noted by Richter and Wilson[3], the economic and mortality impacts of the 1918 Spanish Influenza and 1993 SARs pandemic were radically different from the economic impact of COVID-19 due to fundamental changes in the economy, including increased globalisation, the increase of non-property business interruption covers, and so on.

The second source is unexpected changes in consumer behaviour. Consider the case of COVID-19 cash benefit insurance in Thailand which was sold in 2019 for 300–500 Baht with payouts as high as 500,000 Baht if the insured contracted COVID-19. Thailand initially had a zero-COVID-19 policy but the nation's and individuals' diligence relaxed in 2020 as the less severe Delta and Omicron variants emerged. With lesser severity, consumer behaviour changed from one of extreme caution and self-isolation to looser diligence, with some actually sharing infected masks in order to reap a windfall insurance gain, leading to the bankruptcy of at least four Thai insurance companies.

Caveat 3: More data and advanced models introduce risks of their own

There are four prominent 'new' risks which are introduced or are exacerbated by the use of more data and more complex models.

1. There is an increasing cybersecurity risk and the risk of violating personal data protection rules as more data is collected, used, shared and archived, potentially leading to dissatisfied customers and remediation costs, as well as regulatory fines, which can be mitigated by robust data privacy and information security policies.
2. There is an increasing risk of operational failures and business continuity issues given the increased reliance on global data and computational networks, mitigated by a robust operational risk and business continuity management policy.
3. There is an increasing risk of unconscious bias or regulatory challenges, especially in circumstances where public policy and the economics of underwriting diverge.
4. Simpler models may be more robust, either in predictive power or in formulating an understanding of the situation when faced with dramatic regime change or uncertainty. There is a reason why Occam's razor is a commonly accepted starting point for developing hypothesis and, from the hypothesis, models.

Caveat 4: Behavioural bias and overconfidence

Finally, complex models can generate a 'cult' with professionals more likely to accept the results the more arcane and abstruse the model, even if they contradict common sense; the author has elaborated on this topic in earlier work.[4] In his experience, there is likely to be less focus on the results if they literally are 'rocket science' developed by PhDs (and if the model gives the results that the modeller wants, as was the case for CDO risk and rating models before the 2008 crash).

MITIGATE THE PITFALLS: BALANCING MODEL RISK MANAGEMENT AND COMMON SENSE

These caveats point to potentially severe consequences for financial institutions who blindly follow a 'more is better' approach to data and modelling sophistication. Arguably, some of the risk can be mitigated by following a rigorous model risk management (MRM) framework consisting of five high-level principles (paraphrased from the Bank of England[5], below). These principles are well documented elsewhere and will not be elaborated on more in this paper.

1. *Model identification and model risk classification*: firms should have a definition of a model and resulting scope for MRM, a model inventory and a risk-based tiering to categorise models.
2. *Governance*: firms should have strong governance oversight, policies, accountabilities, culture and a clear statement regarding model risk appetite defined by the Board.
3. *Model development, implementation and use*: firms should have a robust model development process and associated standards covering model implementation, selection and performance. Regular testing of data, assumptions and outcomes should be performed and acted on.
4. *Independent model validation*: firms should have an ongoing validation process that provides independent and effective challenge to models.
5. *Model risk mitigants*: firms should establish and use model risk mitigants for underperforming models.

While these MRM principles are important, from experience I would propose more pragmatic, common sense-based principles. First, put a high emphasis on *common sense* when challenging models, assumptions and results in the context of the model's intended use, stepping back from a blind acceptance that more data or more sophisticated models are automatically better and consciously testing for the caveats outlined earlier.

Secondly, use an *acquire-evaluate-update iterative process*. Ultimately, acquiring data and modelling to get good business results can be viewed as analogue to a child's learning about its uncertain environment, eg an iterative process described by *acquire data* (often through intelligently forming hypothesis, designing and running experiments), *evaluate the results* to refine the economic hypothesis, *update* the model based on the learnings and *repeat* the process. For example, at Allianz Ayudhya, we begin with a hypothesis about a business relevant question, acquire data by designing experiments with control groups (eg experimenting to see if SMS, e-mail, outbound call or agent contact is more effective in generating cross/up-sell revenue for different {product, customer segment} combinations) and updating our hypothesis and propensity models based on the results to improve our business outcomes in the next iteration. Consistent with the acquire-evaluate-update-repeat cycle, Charnes *et al.*[6] suggest the following pragmatic principles:

1. Be parsimonious: start small and add.
2. Avoid mega-models: break the problem into bite-sized bits, divide and conquer.
3. Do not fall in love with data: there is no substitute for careful thought or hypothesis or 'storytelling' about the underlying economic rationale which might have generated the data observed.
4. 'Muddling through': make use of insights, formulate hypothesis, experiment and refine.

CONCLUSION

Returning to the initial question posed by this paper, it is clear from the business use cases that there is often significant value in acquiring and using data and improving financial and risk models. However, that value is tempered by how the large datasets and complex models are used and integrated into a business. Being aware of the caveats discussed in this paper, using common sense, putting in place the appropriate controls and taking an acquire-evaluate-update-repeat iterative approach will go a long way in increasing the value of acquiring data and developing financial risk models.

References

1 Lutkevich, B. and Wigmore, I. (2023) '3V's (Volume, Velocity and Variety)', WhatIs.com, available at https://www.techtarget.com/whatis/definition/3Vs#:~:text=The%203%20V's%20(volume%2C%20velocity,number%20of%20types%20of%20data (accessed 3rd November, 2023).

2 Adobe (2022) 'Types of Analytics Explained — Descriptive, Predictive, Prescriptive, and More', available at https://business.adobe.com/blog/basics/descriptive-predictive-prescriptive-analytics-explained#:~:text=The%20four%20types%20of%20data,in%20the%20past%20(diagnostic) (accessed 3rd November, 2023).

3 Richter, A. and Wilson, T. (2020) 'Covid-19: Implications for Insurer Risk Management and the Insurability of Pandemic Risk', *Geneva Risk Insurance Review*, Vol. 45, No. 2, pp. 171–99.

4 Wilson, T. (2007) 'How I became a Quant', in Lindsay, R. R. and Schacter, B. (eds), 'How I became a Quant', John Wiley & Sons, pp. 95–106.

5 Bank of England (2023) 'Policy Statement PS6/23 — Model Risk Management Principles for Banks', available at https://www.bankofengland.co.uk/prudential-regulation/publication/2023/may/model-risk-management-principles-for-banks?utm_source=Bank+of+England+updates&utm_campaign=f881dbdb03-EMAIL_CAMPAIGN_2023_05_17_08_25&utm_medium=email&utm_term=0_-f881dbdb03-%5BLIST_EMAIL_ID%5D (accessed 3rd November, 2023).

6 Charnes, J. M., Morrice, D. J., Brunner, D. T. and Swain, J. J. (1996) 'Five Simple Principles of Modelling', Proceedings of the 1996 Winter Simulation Conference, available at https://www.informs-sim.org/wsc96papers/104.pdf (accessed 3rd November, 2023).