Predicting quality failures in higher education

Alex Griffiths notes the limitations of a data-driven, risk-based approach for predicting failure

Since 2010 the UK has seen rapid growth in the number of new higher education providers. This growth, aided by reduced barriers to entry to the higher education sector, and concerns over the quality of the new provision it has brought, has been a key driver in successive UK governments pushing for the introduction of a data-driven, risk-based approach to regulating quality in higher education (BIS, 2011; Quality Assessment Review Steering Group, 2015). For the regulator, the Quality Assurance Agency for UK Higher Education (QAA), to prioritize its oversight activity based on freely available performance data has its attractions as high quality providers are allowed to prosper when freed from the burden of unnecessary regulation, and low quality provision is quickly targeted and addressed, and all of this is achieved at a reduced cost to the taxpayer.

A data-driven, risk-based approach, however, relies on one central assumption: that the available data is actually helpful in prioritizing the regulator's activity. Whether or not this is the case has been the focus of an ESRC-funded PhD at King's College London. Our analysis suggests that there is no way to reliably prioritize higher education providers for review despite the wealth of available performance data.

Research design

The research was premised on the fact that we had the outcome of all QAA reviews comparable to today's approach and access to vast amount of historic performance data. This allowed us to investigate whether those providers who were judged 'unsatisfactory' after a review could have been identified in advance using data available at the time. If so, then, in principle, a data-driven, risk-based approach to quality assurance could have been used effectively in the past and our research findings could help inform future risk-based approaches. If it proved impossible to identify high risk providers, even with the benefit of hindsight, our research would suggest that any risk-based approach is unlikely to succeed in the future.

We made use of modern machine-learning techniques to, in effect, try every possible weighted combination of indicators to separately develop the best predictive model for universities, further education colleges and 'alternative' providers. To be as comprehensive as possible we considered not just the indicators in their given form, but also how each provider's performance had changed over time and, where appropriate, standardized indicators by academic year to account for sector-wide shifts in performance over time.

Results

Across all the provider types very few indicators had a strong correlation with the outcome of QAA reviews. Those that did supported the prediction of a small number of 'satisfactory' providers but were of limited use for predicting 'unsatisfactory' providers.

For universities we had 1,700 indicators derived from a wealth of data sources including student surveys, the outcome of previous reviews, complaints raised with the QAA, and staffing, student, research, applications, finance, and overseas activity data.

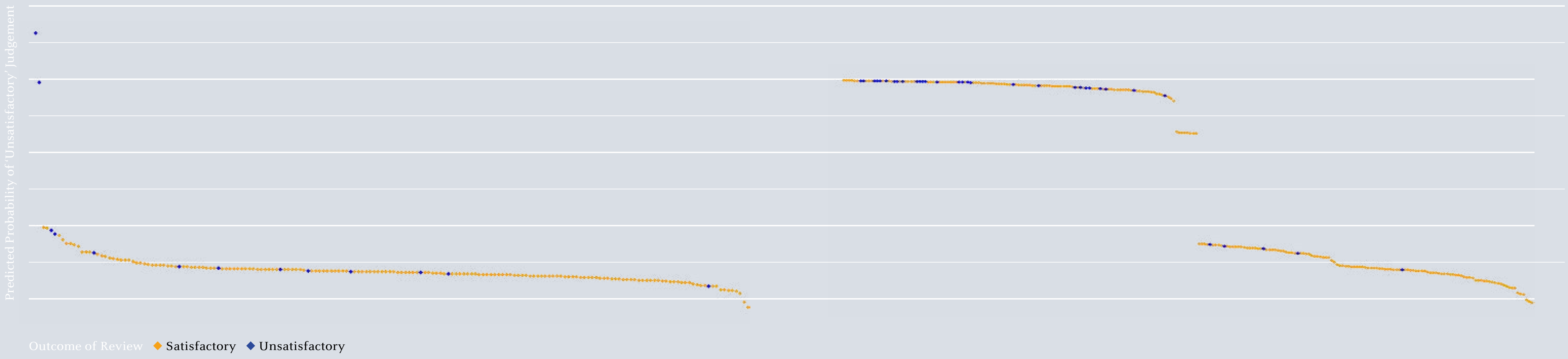
Figure 1 shows the predicted probability of a university being found 'unsatisfactory' prior to the review, ordered from most to least likely, mimicking the order in which the QAA may be expected to prioritize each university, and the subsequent review finding.

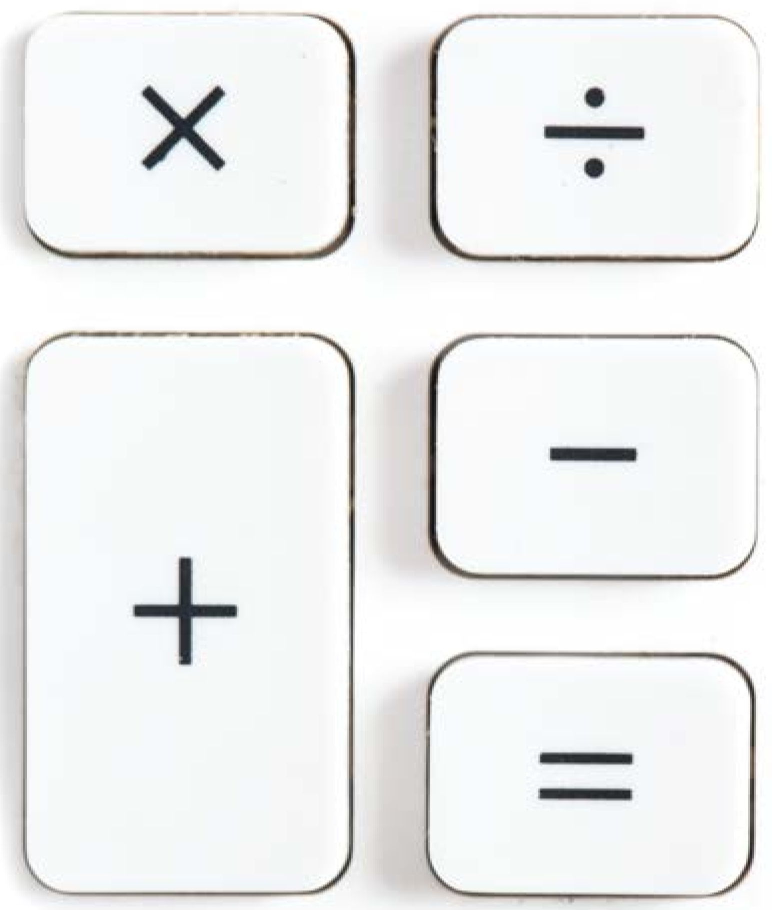
Despite the abundance of data the best model was very poor at predicting the outcome of QAA reviews. Had the QAA carried out their reviews in order of the predicted probabilities, 174 out of the 184 reviews that took place would have been required to discover all 'unsatisfactory' provision and 92.5% of those universities reviewed would have been judged 'satisfactory'. Moreover, with the predicted likelihood of being judged 'unsatisfactory' differing little between universities natural variation in scores would play a large part in the perceived risk posed by each university. Finally, when applied to new data the model produces some questionable results.

The results were similar for further education colleges. The best model required nearly all providers to be prioritized before all of the 'unsatisfactory' provision judgement would have been discovered. However, when the model was tested on new reviews which have taken place since the analysis was conducted, the resulting predictions were worse than chance. The QAA would have be better off doing to the exact opposite of what the model suggested.

Alternative providers offered the greatest promise. There was a clear pattern for younger providers with no prior experience of regulatory reviews and limited funds were significantly more likely to be judged 'unsatisfactory' than more established alternative pro-

Predicted Probability of Receiving an 'Unsatisfactory' Review Judgement: Universities





viders. But, the overwhelming majority of providers would still have had to have been reviewed in order to have discovered all 'unsatisfactory' provision. Reassuringly, the model performed similarly when predicting the outcome of additional reviews.

The most promising finding, true for each of the models, was that the overwhelming majority of the 'unsatisfactory' providers were predicted as being in the 50% of providers most likely to be judged 'unsatisfactory'.

Discussion

The results raise a number of interesting points.

First, regardless of how we define success it is likely any predictive model will disappoint. If it is considered unacceptable to allow any 'unsatisfactory' provision to go undetected then a data-driven, risk-based approach will fail as no model can successfully prioritize all 'unsatisfactory' provision. If it is considered acceptable to allow some 'unsatisfactory' provision to go undetected then the approach is still unlikely to succeed; although the models describe an historical situation satisfactorily, when applied to new data the best models still perform poorly. Either way high quality providers will be prioritized for review, and be unfairly stigmatized as a result, whilst low quality providers will go undetected.

Second, why, despite having the benefit of hindsight and undertaking a comprehensive analysis, could we find no model which could reliably predict the outcome of QAA reviews? There is no shortage of possible explanations: inconsistency in QAA reviewer decisions, concerns over the accuracy of the metrics, the inability of indicators to capture human behaviour, the metrics and the QAA simply measuring different things, or the contested, ambiguous and often changing nature of 'quality' in higher education to mention just some.

Third, if no combination of indicators could reliably inform a risk-based approach in higher education how many other regulators are labouring unknowingly with the same impossible task?

Recent noises from the higher education sector suggests a shift in approach to the interpretation of indicators by a panel of experts familiar with each provider's context (Kimber, 2015; BIS, 2015). How much this undermines the 'rational' and 'objective' prioritization that helped make risk-based approaches attractive to begin with, and perhaps more importantly, whether this leads to an improvement in risk predictions, is yet to be seen. The established literature on the skill of expert judgements does not suggest it will. Hundreds of studies in fields as diverse as medicine, education, finance, and even the forecasting of the future value of Bordeaux wines have consistently shown that the predictions of cheap, simple, rules-based models outperform experts and their unconscious biases (Meehl, 1954; Ashenfelter, 2008; Kahneman, 2011: 234–44).

References

Ashenfelter, O. (2008) Predicting the quality and prices of Bordeaux wine. Economic Journal 118 (529): F174–84.

BIS (2011) Higher Education: students at the heart of the system. Cm 8122. London: Department for Business Innovation & Skills.

BIS (2015) Fulfilling Our Potential: teaching excellence, social mobility and student choice. Cm 9141. London: Department for Business Innovation & Skills.

Kahneman, D. (2011) Thinking, Fast and Slow. New York: Farrer, Strauss and Giroux.

Kimber, I. (2015) Metrics and Quality: do the numbers add up? <http://wonkhe.com/blogs/metrics-and-quality-do-the-numbers-add-up/> Accessed 2 November 2015.

Meehl, P.E. (1954) Clinical Versus Statistical Prediction: a theoretical analysis and a review of the evidence. Minneapolis: University of Minnesota.

Quality Assessment Review Steering Group. (2015) The Future of Quality Assessment in Higher Education. <http://www.hefce.ac.uk/media/hefce/content/whatwedo/learningandteaching/assuringquality/qareview/discussion/QAR_Discussion.pdf> Accessed 15 January 2015.

Alex Griffiths is an ESRC scholarship doctoral researcher at King's College London.