Institution: Imperial College London

Unit of Assessment: 10 Mathematical Sciences

Title of case study: C2 - Improved scorecard evolution methods impacting financial services

1. Summary of the impact (indicative maximum 100 words)

This case study describes impact resulting from research on assessing the performance of credit scoring models conducted by the Consumer Credit / Retail Banking Research Group of the Mathematics Department at Imperial College. The group's work has influenced both high-level industry strategies for developing scoring models, and also low-level performance measures for which such models are developed, refined and evaluated. We describe examples of companies or bodies that have benefitted from improved credit scoring models, including Prescient Models (a US credit scoring company), Experian and the US Office of the Comptroller of Currency. The group has established a very significant reputation for a wide range of commercially valuable work in this area - to the extent that the group received the major *Credit Collections and Risk* industry award for *Contributions to the Credit Industry* in 2012.

2. Underpinning research (indicative maximum 500 words)

The research impact described in this Case Study is concerned with predictive models for guiding decisions concerning individual applicants and users of financial products, such as bank loans, credit cards, car finance, store cards, debit cards, mortgages, student loans, and so on. We focus on *retail* credit - that is, financial services for individual people, not corporations or investors, etc. - and our aim is to model and predict likely behaviour.

A 'scorecard' is a *statistical model* purporting to *measure* someone's riskiness, creditworthiness or other attribute. Such models are used as the basis of loan decisions, to monitor credit card transactions patterns, for fraud detection and for a host of other reasons. Measuring the performance of scorecards lies at the heart of their construction (e.g. parameter estimation to maximise performance), their selection (e.g. which one should be used) and their effectiveness (e.g. is it good enough for purpose). In short, sound evaluation methods are central to this trillion dollar industry's effectiveness and progress.

We describe three of the evaluation areas where we have had significant impact:

(i) **The H-measure**. The most widely used measure of scorecard performance in the UK is the Gini coefficient (GC), which is applied in situations where the aim is to assign people to classes (e.g. good risk or bad risk). The GC is a chance-standardised version of the area under the Receiver Operating Characteristic curve. It is also very widely used in other areas, such as diagnostic medicine, fault detection and signal detection. Currently, this measure is used as the choice of performance measure in around 6,000 papers per year [1]. As the culmination of an extended piece of work, dating from 1999 and continuing to 2009 [2], 2010 [3], 2012 [1] and beyond, we characterised two distinct situations under which the GC may be used. One ignores the classification of other people when assigning a particular individual, and the other takes other classifications into account. We showed that the former usage implies a *fundamental incoherence* in the GC. That is, when used in such situations, *it is equivalent to using different performance measures for different scorecards* - contravening the fundamental performance assessment tenet that the same instrument must be used to measure different scorecards. This contribution therefore identified, and provided a solution to, a deep conceptual problem at the core of all practical scorecard assessment. This solution is a new statistic (the *H-measure*) which overcomes the problem. Public domain code for this is available in the R statistical computing language (http://www.hmeasure.net/).

(ii) **A KS comparative test**. Paralleling the H-measure, the most widely used measure of scorecard performance in the US is the Kolmogorov Smirnov KS test statistic (not the statistical test, *per se*, but the statistic itself). However, we recognised there was no formal statistical test to compare KS statistics: comparisons in the industry had hitherto been based on ranking, or on

informal assessments of the relative size of the statistics. This was a serious shortcoming; it compromised credit-granters' legal obligation to claim that credit-granting decisions are objective, and it risked poor decisions about the choice of scoring model, with adverse implications for both lenders and borrowers. Having identified the problem, to meet the industry needs we developed, described and implemented a test for comparing KS statistics [4].

(iii) **The illusion of progress**. Although scorecards are measuring devices, not classifiers, they are often used as the basis for classifiers by comparing the score with a threshold. More general research on comparative evaluation of classifiers, culminating in [5], demonstrated that there was a pronounced tendency to exaggerate new results, and elucidated the various mechanisms behind this tendency. This contribution is important to banks, since it provides a balancing view to overstated claims of improved performance: it helps them make informed decisions about when new models should be adopted, or when performance claims were (usually accidentally) inflated. [N.B. This work has also had a wide impact in other areas, not least the quantitative algorithmic trading hedge fund industry.]

The Consumer Credit/Retail Banking research group has been located at Imperial College since 1999. The group at Imperial has included over 20 researchers. Key personnel are:

- Prof David Hand, Professor, group leader, 1999-present.
- Dr Niall Adams, PDRA, Lecturer, and now Reader in Statistics, 1999-present.
- Dr Christoforos Anagnostopoulos, PhD student, then lecturer at Imperial, 2006-present.

The research was supported by EPSRC [e.g. G1, G2]. Close collaborations with industry (e.g. Barclaycard, Experian, Fair Isaac, Equifax, Capital One & GMAC) also funded the research [e.g. G3], identified relevant problems, and provided data and other resources.

3. References to the research (* References that best indicate quality of underpinning research)

[1] Hand D.J., Anagnostopoulos C., '*When is the area under the receiver operating characteristic curve an appropriate measure of classifier performance?*', Pattern Recognition Letters, 34, 492-495 (2013). DOI.
[2] *Hand D.J., '*Measuring classifier performance: a coherent alternative to the area under the ROC curve*', Machine Learning, 77, 103-123 (2009). DOI.
[3] *Hand D.J., '*Evaluating diagnostic tests: the area under the ROC curve and the balance of errors*', Statistics in Medicine, 29, 1502-1510 (2010). DOI.
[4] Krzanowski W.J., Hand D.J., '*Testing the difference between two Kolmogorov-Smirnov values in the context of Receiver Operating Characteristic curves*', Journal of Applied Statistics, 38, 437-450 (2011). DOI.
[5] *Hand D.J., '*Classifier Technology and the Illusion of Progress*', Statistical Science, 21, 1-14 (2006). DOI.

Grants:
[G1] EPSRC, EP/C532589/1, 1/10/05/-31/3/08, 'Statistical and machine learning tools for plastic card and other personal banking fraud detection', PI: DJ Hand, CoI: NM Adams, £233,935
[G2] EPSRC, EP/D505380/1,1/3/06-28/2/09, '*Risk Management in the Personal Financial Services Sector*', PI: DJ Hand, Project partners: Fair, Isaac & Company Incorporated, Lloyds Tsb Bank Plc, £540,154
[G3] Link Financial, MATH_P06509, 1/10/06-30/9/09, '*Creating a predictive portfolio review model*', PI:DJ Hand, £83,492

4. Details of the impact (indicative maximum 750 words)

Credit is an important, driving mechanism in the economy. Not giving it when we should has impact on the individual and impact on wider society. The advances in the three evaluation areas described in section 2 benefit consumers and the economy by helping to ensure that consumers are not denied credit when they do indeed qualify and, conversely, preventing people from obtaining credit when they shouldn't qualify.

Since the challenges motivating our research in this area arise from the industry itself,

dissemination of our work and results across the industry is an integral part of our activity. We achieve this dissemination in a number of ways, such as formal consultancy projects, industry-funded PhD studentships (sometimes part time, with company employees), postdocs, invited presentations to corporations, and through industry conferences (where we are regularly invited to give keynote presentations). Examples of bodies which have funded our work include GMAC, (evaluating scorecards), HBOS (pattern discovery in retail banking data), British Credit Trust (developing new scorecards), Fair Isaac (evaluating scorecards), Capital One (evaluating scorecards), Goldman Sachs (fraud detection), and many others. We have worked with most of the major industry players, and many minor ones.

Turning to each of the three areas of research described in Section 2 we now describe the impact:

(i) **The H-measure:** These matters were first presented (by Hand) to the industry at the *Henry Stewart Conference on Predictive Analytics* (an important conference for industry users of marketing analytics) in December 2008. Since then Hand has been presented this work in many commercial and industrial contexts (as well as more academic meetings), for example, *Credit Scoring and Control XI* (August 2009, the premier conference on retail finance, with 90% industrial participants), *IMA Conference on Mathematics and its Applications* (March 2011), an invited seminar to *Opera Solutions* (November 2011, a leading data analytics company), a two-hour keynote presentation at the Capital One *Allstat and Quants* conference (October 2013), and many many others. The H-measure is being increasingly adopted in the retail credit industry, as a performance measure which overcomes the incoherence problem of the current industry standard Gini and KS measures. The CEO of a leading US credit modelling company, Prescient Models, writes: "*Over the last decade I have found the research papers from the team at Imperial College to be very valuable. For the topics of adverse selection, reject inference, the proper use of statistics like KS and Gini, and survival models, I have read several excellent papers that have assisted in my product development and general understanding*" [A]. He also notes that paper [2] "*has alerted practitioners to how they can make more useful models instead of simply chasing improvements in arbitrary measures*" [B].

(ii) **A KS comparative test:** This new test [4] was first described, by Hand, to the industry at the *Credit Scoring and Control XII* meeting (August, 2011). Of this work, the Director of the Credit Risk Analysis Division of a US banking regulator, the US Office of the Comptroller of Currency, commented in 2012: "*The statistical test you outlined in your paper is exactly what is needed to add statistical rigor to the decision process. It is a test we will recommend banks adopt as part of their model selection and validation process*" [C]. In 2013, the CEO of the US credit modelling company, Prescient Models, noted that the work in [4] provided "*further weight and clarity"* to the issue, allowing *"models to be judged realistically"* [B]. In praising the robust K S comparative test in [4], he commented that "[s]*ubstituting one model for another in pursuit of spurious accuracy can be a multi-million dollar waste of money and distraction*" [B].

(iii) **The illusion of progress:** The illusion of progress work is fundamental, with wide applications, and can be considered a cautionary tale for the industry, warning it against inflated claims of enhanced performance for complicated classifiers. The Director of the Credit Risk Analysis Division of a US banking regulator, the US Office of the Comptroller of Currency, wrote about the work described in [5]: "*I would also like to note the contribution your paper [5] has had on our discussions internally and with modelers at the larger institutions we supervise ... Your insight on this issue has helped us solidify our thoughts on this issue and develop methods of assessing the process banks use to develop and implement their new models. We have used your paper as a "discussion paper" for an internal workshop on modeling methods after the crash, and as a recommended research paper for modelers at large and mid-size banks...*" [C]. Similarly, "*This article has had a direct impact on the creation of models in retail lending with some of its terminology quickly becoming part of standard conversation within the industry. It has provided the needed support for analysts to avoid wasting time on spurious improvements, thereby saving moving and allowing analysts to investigation deeper issues"* [B].

Further impact

In addition to the specific impact that has arisen from our research on the H-measure, KS

comparative test and illusion of progress, the more general impact made by the statistics group is demonstrated by the following statements from people we have worked with in the financial industry, and the major industry award citation given below:

- One of the leaders of the industry, a former CEO of Experian (Experian is a FTSE100 company, and a leading international provider of credit information) writes "*You and your team's research has helped enormously in the development of credit scoring and the credit industry*" [D].
- This general point is also made by an employee of one of world's largest credit scoring organisations, Fair Isaac, who says "*it is more valuable than ever to hear your voice which cuts straight through to the scientific underpinnings of predictive modelling and classification, while at the same time being concerned with the important issues of practicality and usefulness of the resulting models*" and "*We continue to look very much forward to your highly relevant contributions in the field of credit scoring and statistics*" [E].

Finally, in 2012, the Consumer Credit Research Group at Imperial was awarded the Credit Collections and Risk (CCR) industry award for Contributions to the Industry. This is the first time this, or indeed any of the CCR awards, has been made to an academic unit. The award was presented by Gary Brooks, Group Credit Manager at Hitachi Europe, and the citation read: "*... our winners have contributed significantly to improving decision-making strategies for the retail credit industry. They have worked with regulators, the banking and finance sectors and scoring and ratings agencies worldwide in a wide range of areas to improve scoring"* [F]. Prof Hand has also been asked to join the Editorial Advisory Board of *Credit Collections and Risk*, the leading industry magazine in the area.

5. Sources to corroborate the impact (indicative maximum of 10 references)

[A] Letter from CEO, Prescient Models LLC, 2/1/2012 (letter available from Imperial College on request)
[B] Letter from CEO, Prescient Models LLC, 4/4/2013 (letter available from Imperial College on request)
[C] Letter from Director, Credit Risk Analysis Division, US Office of the Comptroller of Currency, 9/01/2012 [The Comptroller of Currency is the US Federal Agency responsible for chartering, regulating, and supervising all national banks and the federal branches and agencies of foreign banks] (letter available from Imperial College on request)
[D] Letter from Former CEO of Experian, now PDG of Scoresoft, 25/10/2011 (letter available from Imperial College on request)
[E] Letter from Analytic Science-Senior Director, FICO Research, FICO (Fair Isaac Corporation), one of the world's biggest players in the area,16/10/13 (letter available from Imperial College on request)
[F] Contribution To The Industry Credit Excellence Award 2012, Winner: Professor David Hand and the Consumer Credit Research Group at Imperial College, http://www.ccr-interactive.co.uk/index.php?option=com_content&task=view&id=1416&Itemid=117 (Archived at https://www.imperial.ac.uk/ref/webarchive/tkf on 21/5/13)