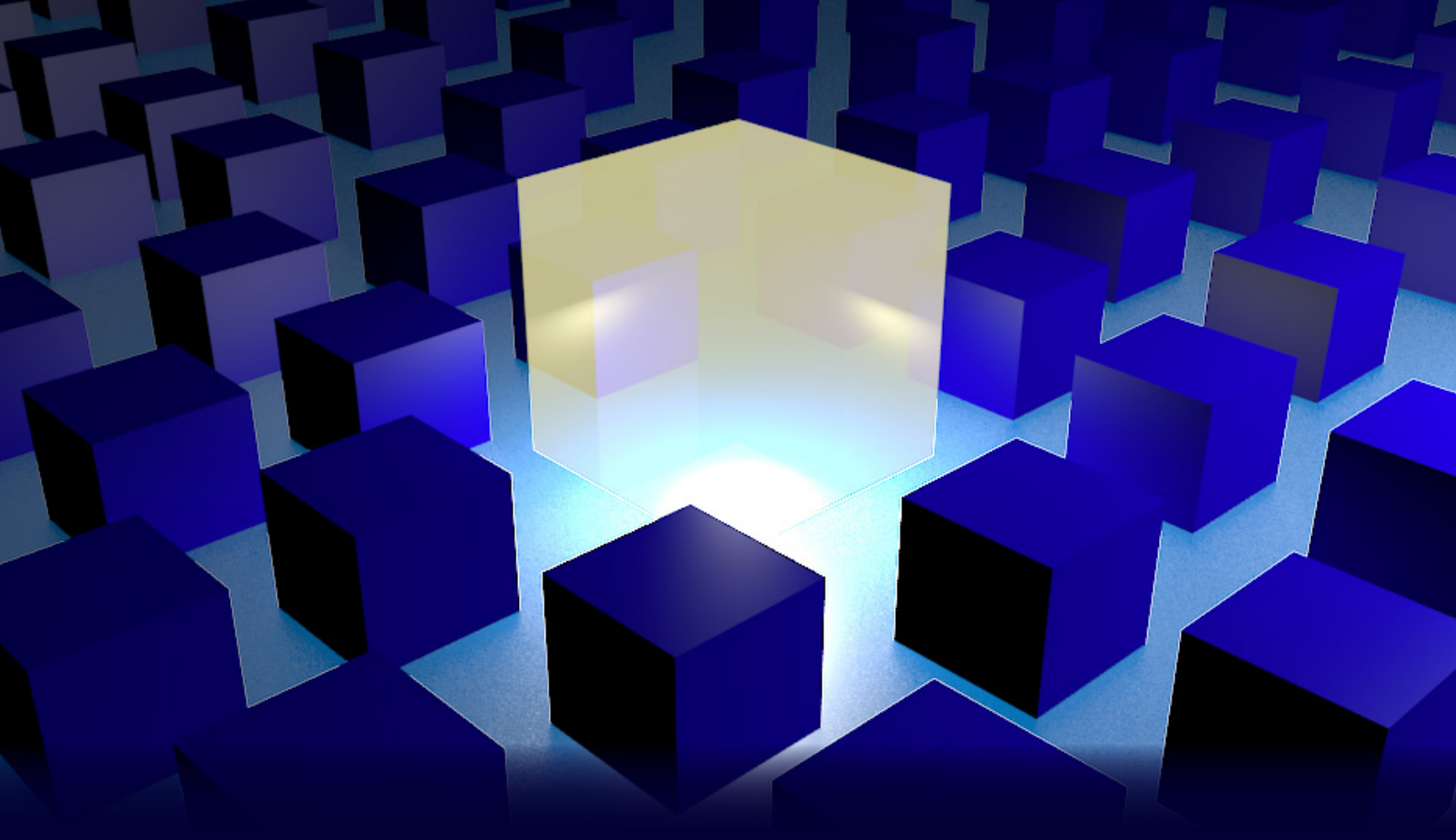

Deep learning and its interpretability in retail banking:

For neural network black boxes, ‘interpretable’ is the new black

Riskcare

CREDIT RISK ASSESSMENT FOR RETAIL BANKING AND DEEP LEARNING

S. CAENAZZO, D. KNIGHT AND K. PONOMAREVA

The Retail Banking industry is primarily concerned with the distribution of financial services to private individuals. From the bank's perspective, the range of services on offer varies from almost credit-risk-free – such as current accounts – to potentially high-risk products that include credit cards, loans and mortgages.

One of the most popular techniques used to assess and manage the credit risk of potential applicants is the so-called credit scorecard, usually obtained via techniques such as logistic regression.

A typical scorecard takes a number of inputs from the credit applicant, such as income and usage history of previous credit products. Each of these inputs is transformed into a score, according to certain rules. The final step is to obtain an aggregate score by summing the individual input scores – if the applicant's score is found to be above a certain threshold, representative of the lender's risk appetite, the application is accepted.

The development of the scorecard's rules and threshold is only possible via analysis of relatively large cohorts of previous customers' data and debt repayment behaviour.

Credit scorecards have proven very useful in this context as they make effective use of relatively coarse data (e.g. account statement and/or income data, which could be annual or semi-annual) and provide an intuitive, visual explanation of which factors contributed the most towards a certain score.

At the same time, it is possible to provide feedback to a rejected applicant regarding the most likely reasons for the rejection; individual scores that are particularly weak are likely to correspond to the factors that are implicitly categorising the applicant as high-risk.

In recent years the retail banking industry has started to see a significant increase in the volume of data available for analysis. There are several sources of this data; social networks, mobile application usage statistics and even government-backed data frameworks, such as Open Banking (www.openbanking.org.uk) – which was rolled out in the UK in January 2018. This latter initiative allows private individuals and SMEs to share their transaction-level account information on a voluntary basis, enabling participating banks and lending institutions (especially smaller ones that may not have enough data to conduct risk analyses) to offer better-tailored credit products to whoever decides to share their data.

This increase in data volume presents both advantages and challenges. On one side, a large dataset is key to generating accurate statistical analysis – generally speaking, more data is always desirable. Conversely, once a dataset becomes too large it may require a paradigm shift in the tools and techniques used to analyse it.

With regards to the credit scorecard approach mentioned previously, an effective scorecard takes a limited number of inputs and this can be a serious limitation when analysing highly-granular datasets. A possible solution is to perform data aggregation upstream to make standardised scoring viable, but at the cost of potential information loss in the process.

Where this is the case, more sophisticated deep learning (DL) models can be utilised to take advantage of their specific capabilities in analysing certain types of granular data. In the past decade there has been a surge in remarkable results attained by DL algorithms – especially neural networks (NNs) – in various pattern recognition applications. These vary from image classification and voice recognition to forecasting of interest rate dynamics.

THE INTERPRETABILITY PROBLEM

The widespread use of advanced DL models in sensitive fields such as medicine or consumer finance has been hindered by a fundamental lack of human interpretability regarding the outputs of such advanced models.

Simpler techniques (e.g. linear or logistic regressions) yield outcomes, which are deterministic in nature and follow mechanics that are well understood and controlled by model developers and analysts.

Neural networks, however, have a large number of hidden layers, the exact roles of which are not easily interpreted. For example, given a picture of a cat, the analyst may be left wondering why the model decided correctly to classify it as a cat. Or worse, if the model unexpectedly classifies it as a dog, it may be unclear why the model got it wrong. Historically, obtaining answers to these questions in the context of neural networks was extremely arduous, if not impossible.

Before we look at potential solutions to this issue, why is human interpretability essential in a discipline such as credit risk estimation? There are several reasons, but let's focus on the two major ones.

Firstly, customers have a fundamental right to be fairly assessed (see, for example, the white papers and directives from the World Economic Forum[1] and the Basel Committee on Banking Supervision[2]). This means that any lender's decision on a credit application must be explainable.

This applies to rejection and acceptance decisions alike: in the first case, the customer has the right to know if any discrimination (e.g. decisions based on ethnicity, gender, etc.) took place, whilst in the second case the customer should be reassured that the bank did not blindly accept a credit application that is likely to end up in heavy financial distress for the applicant – the so-called concept of credit affordability.

Secondly, to be credible, risk management must go through layers of direct human responsibility. It is not acceptable for risk managers or underwriters not to understand the mechanics of the models they are using. It casts a substantial shadow of mistrust over the analysis results, as there is no real way of verifying them and comment on their adherence to reality. Moreover, it also significantly hinders any attempt at developing sound business and/or risk management strategies; every high-level management and/or strategic decisions must be endowed with in-depth information and knowledge about the processes involved. If the outcomes of some models in the process chain cannot be understood and interpreted, any decisions taken on their basis will be inevitably ill-informed.

These considerations have been further echoed in a recent paper by the House of Lords' Select Committee on Artificial Intelligence[3]. The paper explicitly addresses the need for intelligible artificial intelligence (AI) systems and, in particular, mentions the recommendation from experts of the University of Edinburgh to have a high degree of intelligibility for AI/DL systems in certain kinds of financial products and services – among them personal loans and insurance.

MANAGING THE INTERPRETABILITY PROBLEM

Explainability does not require a new category of model, but rather ensuring that outputs from existing models are sufficient to provide understanding – both at a technical level for the developer and for the end user. For existing models, the ability to explain a model might be mapped against the accuracy of the outcome with respect to the 'ground truth' (i.e. the phenomenon we wish to model) and used to generate appropriate metrics for model quality.

'To be credible, risk management must go through layers of direct human responsibility. It is not acceptable for risk managers or underwriters not to understand the mechanics of the models they are using.'

Measuring the performance of different models, DL might score highly for predictive accuracy but low for explainability in comparison to decision trees – which score lower on accuracy but are comparatively easy to explain.

Many existing systems within large organisations are already at a level of complexity beyond the immediate ability of a single person to understand. With the addition of AI, this challenge will only grow.

Even in a relatively explainable model, the outputs may still be too complicated to explain easily and therefore it will be necessary to provide user interfaces and effective data representation techniques to allow additional insight. This will require new skills and creative techniques to illuminate the biases and decision logic embedded within these models. This will range from summarised reporting dashboards, right down to deep explanation and traceability of the decision process.

Ultimately, regulators will have to provide guidance on the fair use of data, allowing organisations to ensure their models are fit for purpose.

[1] *How to Prevent Discriminatory Outcomes in Deep Learning*, World Economic Forum White Paper - Global Future Council on Human Rights 2016-2018.
[2] *Sound Practices - Implications of fintech developments for banks and bank supervisors*, BIS BCBS white paper.
[3] *AI in the UK: ready, willing and able?*, House of Lords Select Committee on Artificial Intelligence, Report of Session 2017-2019, April 2018.

THE SOLUTIONS

Recently, there has been significant work attempting to explain and interpret outputs from neutral networks. The problem has been tackled from several different angles:

RELEVANCE ANALYSIS

How much of the output (e.g. a probability of default) is directly attributable to a given input variable (e.g. the applicant's annual income)?.

MORE DETAILS: As complex as they may be, neural networks are still chains of data manipulation processes – nothing stops us from picturing this as a single process whereby we obtain an output (e.g. a probability of default) as a function of some inputs (e.g. the data of the credit applicant; annual income, past repayment behaviour, etc.).

The final step is for us to define a relevance metric to be calculated for each input variable; if we design this metric such that the sum of the relevance metrics for each input is equal to the final output, we will have a direct way of assessing the role each input variable played to reach the output.

In a credit risk example, we could be able to answer questions such as: 'Does this model give more priority to current income or previous card usage when accepting/rejecting clients?'. Relevance analysis is deceptively simple as a concept (its mathematical roots are not more involved than high-school-level calculus), but it is achieving impressive results in the explanation of various types of neural network.

SENSITIVITY ANALYSIS

How much does the output change according to a (small) change in a given input variable?

MORE DETAILS: Sensitivity analysis, in some form or another, is pervasive in the financial industry: from delta-hedging exercises to Greeks-based Value-at-Risk, sensitivities are a mainstay analytical tool.

The underlying concept is very similar – whether we are considering neural networks or other approaches – and revolves around the analysis of variations in the output conditional to variations in input variables. The normalised output variations or sensitivities can provide very useful information about the system: is there any variable that, if changed even by a small amount, causes a completely different output – perhaps signalling an instability of the model? Is there instead a variable that does not seem to influence the final result, even under massive deviations, and could therefore be dropped from the analysis?

A systematic and comprehensive analysis of these behaviours can be an invaluable tool for model developers and users as it allows detection of potential biases and/or design flaws in the model. Note that sensitivity analysis differs from relevance analysis as it specifically addresses the role of variations in the inputs, rather than the inputs themselves.

NEURAL ACTIVITY ANALYSIS

Which neural paths are most activated by a given input variable?

MORE DETAILS: Neural networks have been historically developed as mathematical representations of the human nervous system – as such, the anthropomorphisation of NNs is not accidental (we often refer to 'decisions' and 'thinking' by NNs, after all).

Just like the human nervous system, NNs react to stimuli (i.e. the input variables) by activating neurons (also called nodes) that subsequently manipulate and transmit the stimuli data to other neurons.

As is the case with the human nervous system, there is a lot to learn by just observing which connections between neurons are most utilised by the network in the presence of certain stimuli. For example, we may concentrate our attention on those areas of the network that are utilised the most: close inspection of their activation patterns will most likely reveal patterns in the data considered significant by the model.

On the other hand, areas of the network that are not activated often may signal an inefficiency of the network, perhaps requiring some review in its design.

EXPLAINABLE ARTIFICIAL INTELLIGENCE (XAI):	
A human cannot read into the mind of an AI… but another AI can!	**MORE DETAILS:** Perhaps the most recent among the methods shown here, Explainable Artificial Intelligence (XAI) revolves around building ancillary NN/AI models with the sole purpose of generating insights into the inner gearing of a target NN/AI process. The nature of such insights may greatly vary depending on the situation: a great example of a successful XAI setup was shown by the joint research between DeepMind, Moorfields Eye Hospital and University College London, where a Deep Learning model has been trained to diagnose eye diseases from the analyses of retinal scans and a second Deep Learning model provides explanations and recommendations on the diagnosis.

All of these techniques have been originally developed in traditional contexts for NNs, primarily in image recognition/classification problems, but they have soon found applications in other fields. Our opinion is that these methods could be particularly useful in the financial domain as well.

WHAT IS THERE TO GAIN FOR FINANCIAL INSTITUTIONS?

As mentioned previously, a historical hurdle around the use of NNs has precisely been the lack of human insight into the mechanics of the underlying algorithms. In light of these innovative developments, it seems like this issue could be finally overcome. A combination of the above techniques may be able to provide complementary insights into these traditionally opaque systems and yield several benefits to the banking industry:

- **LARGE INSTITUTIONS WOULD NOT BE IN THE DARK WHEN IT COMES TO NEURAL NETWORKS.** They could be confident in their ability to explain outputs to business stakeholders, and pro-actively use the model insights (e.g. trends in portfolio credit risks, need for additional analysis of specific variables, etc.) to manage risks and control business performance on the basis of informed decisions;
- **ABILITY TO COMPLY WITH RIGOROUS REGULATORY STANDARDS.** Proper vetting would be feasible from a number of model risk perspectives: from impact analysis of input data up to model performance/ soundness analysis, the previously mentioned interpretability tools may offer the necessary drill-down capability required by regulatory bodies and model risk best practices.
- **SMALLER INSTITUTIONS THAT ALREADY USE DL AND NNS COULD IMPROVE COMPETITIVENESS**. The power of a disruptive technology is constrained by the ability of its wielder to fully understand its nuances. Several FinTechs and start-ups already use NN-based models for a range of banking applications, from highly-specific credit scoring applications to data anomaly and fraud detection. Enhancing the understanding of such models could greatly help the development of even more innovative, efficient and competitive products and technologies.
- **TRANSPARENCY TO THE OUTSIDE WORLD COULD BE INCREASED.** As we have seen, several governing bodies are concerned about the lack of transparency of these models and their impact on social structures – in the financial space, this concern is very much specific to retail banking as it is the closest point of contact between banks and private individual customers. The ability to interpret NN outcomes could bridge the gap between technology developers and final users. This would ultimately culminate in the capability to reassure customers (and every other entity impacted by the use of these models) on the overall validity of the models and the absence of undue biases in them.

CONCLUSION

The interpretation of neural network-based models is an important hurdle when it comes to the widespread adoption of such models in retail banking. Regulators increasingly demand fairness, transparency and equality and the opaque nature of these models can make this a difficult issue.

A number of emerging methods are looking to address this problem, and promise a number of significant benefits. We have primarily discussed retail banking here, but there is no doubt that these interpretability tools could prove invaluable right across the sector.

Our opinion is that deep learning and finance are drawing much closer – the array of analytical methods now available may well be the long-awaited breakthrough that the industry requires to see inside the 'black box' of neural networks. After being largely restricted based on interpretability grounds, banks and financial institutions may finally be able to consider the introduction of NN-based models on a much broader scale

'The interpretation of neural network-based models is an important hurdle when it comes to the widespread adoption of such models in retail banking. Regulators increasingly demand fairness, transparency and equality and the opaque nature of these models can make this a difficult issue.'

ABOUT RISKCARE

Riskcare is a financial services consultancy and outsourcing company with offices in London, New York and Sydney.

Over the past 23 years we have built up experience and knowledge that sets us apart in delivering advanced, complex and transformational change to the capital markets industry.

We service a broad range of clients, including investment banks, institutional investment companies, hedge funds, exchanges, commodities trading houses and insurance corporations.

STAY CONNECTED

LinkedIn

Connect with us:
Riskcare company profile

LONDON
Karen Line

Tel: +44 (0)20 7614 3600
DDI: +44 (0)20 7614 3742
sales_eu@riskcare.com

25 Worship Street
London, EC2A 2DX

NEW YORK
Ian Rebello

Tel: +1 646 625 5415
DDI: +1 646 625 5422
sales_us@riskcare.com

1431 Broadway
New York, NY 10018