BANK OF ENGLAND

FinTech Proof of Concept

Digital Reasoning – big data text analytics

Background

The Bank's Prudential Regulation Authority (PRA) is the United Kingdom's prudential regulator of deposit-takers, insurers and major investment firms. As part of its remit, the PRA supervises over 500 insurance firms including general insurers, life insurers, friendly societies and the London insurance market. These PRA-regulated firms produce and publish a large amount of weakly-structured textual information. The quantity and variety of such data is expanding rapidly, and the Bank is keen to explore where new technology can support and improve existing analytical techniques in extracting policy-relevant information. As such, a tool that enables supervisors to search through a corpus of documents for sections relevant to a given theme, and which automatically flags new information of regulatory interest for review, could potentially support the Bank's supervisory approach.

The Proof of Concept

In the fourth cohort of the FinTech Accelerator, the Bank worked with Digital Reasoning on a Proof of Concept (PoC) to demonstrate the analytical value of a machine learning tool that could ingest and analyse large quantities of weakly-structured text data and detect patterns, anomalies and themes. For the purpose of this PoC, we explored the ability of Digital Reasoning's Cognition tool, which uses supervised learning techniques to map relevant sections of text automatically into categories based on regulatory interest.

To demonstrate the capabilities of Cognition, Digital Reasoning ingested data consisting of a sample of publically available information of potential regulatory interest to the Bank's insurance supervisory teams. In total, approximately 10,000 data points were loaded into the tool in order to train it with input from experienced analysts.

The Cognition tool enabled users to build 'annotation models' based on particular themes or 'labels'. For each model, users applied labels to excerpts of text selected by the software. For the purpose of

this PoC, we trained models in line with the PRA's supervisory risk framework, which considers the overall risk context for insurance firms in terms of 'external context' and 'business risk'.[1]

During the training process, Cognition used supervised machine learning techniques to adjust the model iteratively in order to improve understanding of which sections of text might be considered as examples of business risk and external context respectively. The tool also features a 'cross-validation' feature, which provides information on the accuracy of each model as it is trained by the user using an increasing number of examples. After a number of iterations of user feedback, based on the cross-validation results presented by the tool, we were successfully able to use the model to categorise instances of business risk and external context in the data.

The tool also included a multi-labelling feature, allowing users to train multiple models at the same time (for example to identify text relating to both business risk and external context within the same data set). We found that this enabled training to be carried out more efficiently. However, as the number of labels for the model increased, so did the minimum number of annotations required to train a model. In addition, the current configuration of the software did not allow application of multiple labels to the same excerpt, as they needed to be considered mutually exclusive.

We worked with Digital Reasoning to design a dashboard which could be used to summarise and navigate through the outputs of Cognition. We found this to be a useful tool for filtering results pertaining to particular firms and time periods, providing a snapshot of valuable risk related information.

Reflections and next steps

We found that Digital Reasoning's Cognition tool could usefully categorise sections of text in the data based on themes of regulatory interest. In addition, we found that the underlying model became more accurate as the model was trained over time. We found the user interface to be intuitive, which enabled analysts to train models quickly with little formal training on the tool.

This PoC demonstrated that machine learning techniques can provide useful insights into textual data sources. The dashboard demonstrated in concept how the outputs of machine learning categorisation can be presented in a practical way for regulatory use despite some technical limitations. The Bank will continue to evaluate the Cognition tool over a limited trial period.

[1] https://www.bankofengland.co.uk/-/media/boe/files/prudential-regulation/publication/the-pras-approach-to-insurance-supervision-march-2016.pdf?la=en&hash=E1DA751CDCDADF2651EFBEC94E0814C6ACD7D798