BRIEFING PAPER SEPTEMBER 2015

LEARNING FROM THE CIVIL SOCIETY CHALLENGE FUND: PREDICTIVE MODELLING

KEY INSIGHTS

This paper tests a new methodological approach to analysing data in order to predict outcomes, which could be used by future fund managers and monitoring and evaluation specialists. It analyses the scores assigned to proposals, annual reports and project completion reports by the Fund Manager. The following learning has arisen from this new approach:

1. Decision Tree models can be quickly developed and evaluated by using open source software and a modest amount of online tuition.
2. The use of case-based knowledge held by staff is important at the planning and conclusion stages of prediction modelling.
3. The quality of the models is dependent on the quality and relevance of the underlying data.

THE CHALLENGE AND OPPORTUNITY

The ability to predict project outcomes should be of interest both to funders and implementers of development projects.

"Prediction is very difficult, especially if it's about the future." Nils Bohr

If enough programme data has been collected it is possible to identify patterns in the data. Rules can be

The **CIVIL SOCIETY CHALLENGE FUND** (CSCF) was a demand-led fund which aimed to enable poor and marginalised people to have a voice on issues that affect them and to be included in local and national decision making forums. Running from 2000 to 2015, it supported 526 projects in Africa, Asia, the Americas and the Middle East, each with a grant of up to £500,000 and running for 3 to 5 years.

This learning brief is one of six, prepared upon completion of the CSCF, focusing on key areas of best practice within the fund. These briefs aim to share learning with practitioners and civil society learning networks, and help inform future fund management in DFID and beyond.

"Ninety per cent of problems have already been solved in some other field."
Tony McCaffrey

created from those patterns and tested on the same data to see whether they prove true. The more accurate the rules, the more confident one can be that programmes applying those rules to their design would also be successful.

Using data over the last 5 years, the Fund Manager has accumulated detailed information on 60 CSCF projects. In a data set where each project may be described using more than 170 project attributes, there are many millions of possible combinations that might be good predictors of the final outcome of these projects.

A theory-led inquiry will focus on a very small sub-set of these possiblities: additional strategies are needed to systematically and comprehensively search for configurations of project attributes that are associated with the project outcomes of most interest.

Development projects rarely work with very large data sets, but these are much more common in the scientific community and large businesses. In these fields, an array of data mining tools have been developed to help identify patterns in data sets that may be meaningful. In one approach, the use of Decision Tree algorithms has potential value in the evaluation and analysis of development results.

They have three advantages:

1. The results that are generated by the algorithm are easy to read. The diagram below shows a Decision Tree model built using CSCF project data. Each "leaf" in this inverted tree is a type of outcome: 1 represents "more effective" projects and 2 represents "less effective" projects. The nodes in the tree represent different project attributes that may be present or absent. Reading down the tree from the top we can see that where capacity building of local government is present (=1.0), but there is no capacity building of private sector organisations (=0.0), but there is capacity building of end beneficiaries (=1.0) then there are more successful projects (square =1.0). The coloured band in the square tells us how many actual projects fit this rule. Almost all do (blue), but one does not (red).

2. The results generated by Decision Tree algorithms show that they are sufficiently representative of reality, but not too complex. As Figure 1 shows, there is more than one way of achieving "more effective" projects. Successful outcomes can be the result of

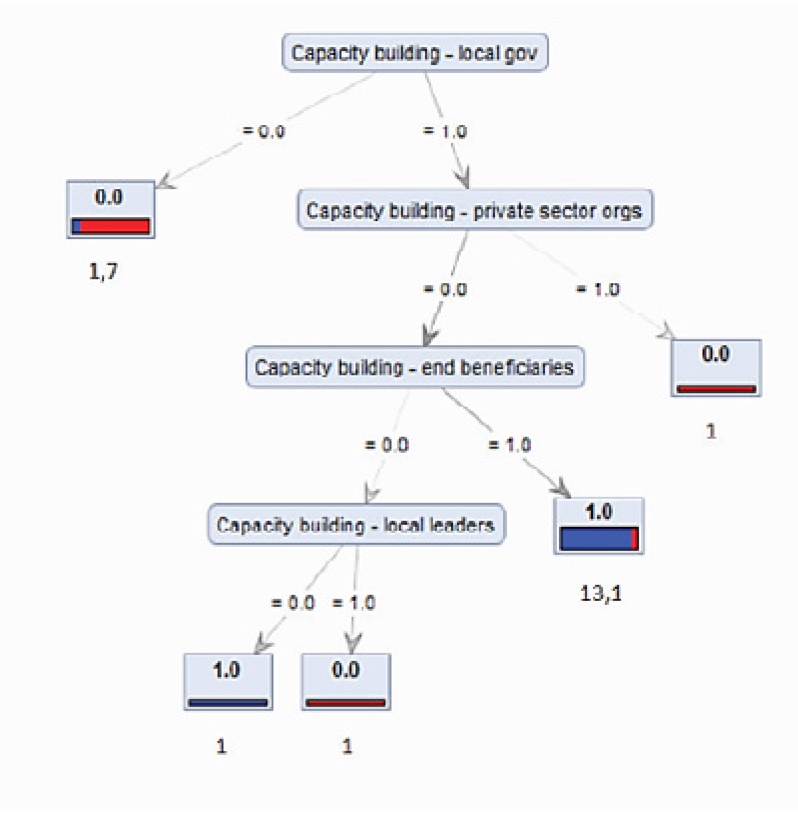

Fig 1. Tree model

		Outcome is observed:	
		Present	Absent
Model expects outcome to be:	Present	14	1
	Absent	1	9

Table 1. Truth table

combinations of project attributes, not just single factors. The causes of less successful outcomes may not be just the absence of the causes of success, but quite different factors altogether.

3 There are many ways of measuring the performance of these kinds of predictive models, all of which make use of numbers of cases in a truth table. Table 1 summarises the tree model in Figure 1. It has a high level of "accuracy": (14+9)/(14+1+1+9) =92%.

THE APPLICATION TO CSCF PROJECT DATA

The available data was collated and structured into one Excel file with 53 project attributes and 21 project outcomes. Data that was available on 35 projects was used as a training data set to develop an initial set of 26 predictive models in the form of Decison Trees. The simplest predictive rules in these models were then tested against data that later became available on the remaining 25 projects. All but two of the 39 rules performed better than chance. From these, the best performing example models (representing 50% of the 39 rules) were kept and then screened by the Fund Manager for any inconsistency with their own case-based knowledge.

KEY INSIGHTS INTO THE WORK ON PREDICTIVE MODELLING

1 The development of predictive models can be done quickly and easily

Individual predictive models can be generated and evaluated within approximately 30 minutes. In addition, preparation time needs to be invested in collating and cleaning existing data, and setting up a modular analytic process within the software (RapidMiner).

It is then possible to simplify the most useful branches of content of Decision Tree models into relatively simple "IF…AND… THEN…" types of rules.

For example, projects which had 'no issues' in their recent risk rating and did engage in policy engagement, and did build the capacity of local government, but did not build the capacity of private sector organisations achieved higher effectiveness scores. This configuration covered 71% of the projects with these outcomes, and predicted their outcome with 100% accuracy.

2 Case-based knowledge is also important

At the planning stage choices need to be made about what types of project attributes and outcomes should be included in the modelling, and what broad types of relationships between these should be analysed (for example, between initial project proposal appraisal ratings and subsequent project performance ratings).

Once the modelling results have been generated these need to be screened to identify rules that contradict existing expectations. This is where examination of exemplar cases is essential - to identify if there are underlying causal mechanisms at work or if the association has no causal basis.

For example, some CSCF fund management staff felt that this prediction rule contradicted their current knowledge: "Projects with neither an objective of innovative service delivery nor engagement with national decision making achieve some or significant

EXAMPLES OF PREDICTION RULES

Projects which had "no issues" in their recent risk status and "minor or no issues" in their finance rating achieved higher effectiveness scores.

The attributes in this rule fitted 81% of all projects and the prediction was accurate for 78% of those cases. This appears to be significant, given that a random choice would have 53% accuracy.

Projects which did not address national decision making were associated with few or no achievements in change in discourse.

The attributes in the rule fitted 64% of all projects and the prediction was accurate for 79% of those cases.

improvement in innovative service delivery."

Of the initial set of prediction rules, 38% were assessed as confirming existing knowledge. Fund Manager views were divided or undecided on the remaining 62%, suggesting further screening was needed.

3 Data quality matters

Of the 23 Decision Tree models developed in the first stage of analysis, 9 were judged as weak or inadequate. The main cause was insufficient diversity of rating scores, notably for the outcomes of "sustainability" and "project performance", where one score was dominant. In other cases, there was insufficient diversity of types of project attributes to generate a good predictive model. For example, there were were only three project partner attributes, and within these there was little differentiation of rating scores. This demonstrates the importance of thinking carefully about likely future data use when designing scoring systems to analyse projects, proposals and results.

RECOMMENDATIONS AND CAVEATS

1. Data that is collected on project outcomes should be carefully designed to capture the range of performances that exist, in order to be able to develop useful predictive models. The same applies to data on project attributes that are potential predictors of project outcomes.
2. If data is being collected on projects on an annual basis, as was the case with the CSCF, then it should be possible to update predictive models on an annual basis, and use this knowledge to inform the management of grant making "in process" rather than only at the closure of the Fund.
3. Just because two elements are associated in the model does not necessarily mean that one element is causing another. Finding associations through the use of algorithms like Decision Trees helps us to identify what areas we should then investigate in detail through careful within-case inquiries. More work still needs to be done on the design of appropriate screening strategies, once a set of prediction rules has been generated.
4. Decision Trees are useful models, but represent very simplified views of the world. Each branch in a Decision Tree represents a possible causal configuration but not a casual pathway. There is no implication that the capacity building activities in each branch in Figure 1 have to take place in the order they appear in the branch.
5. Packages like RapidMiner are one option. There may be alternative means of doing similar but simpler kinds of analysis. The Fund Manager used an excel application for analysis of CSCF data and continues to be developed beyond the the lifetime of the CSCF for other projects/programmes - or words to that effect.

CONCLUSION

Although this is a relatively small data set, it does show potential for wider application by Fund Managers when screening proposals to predict project results.

Predictive modelling relies on highly qualified technical Fund Managers applying scoring systems consistently across a portfolio of projects. Trends in data can only be achieved if scoring systems do not change. The results of predictive models should always be checked against what is known about individual cases they refer to.

REFERENCES AND LINKS

Tony McCaffrey, quoted in article 'Eureka machines', New Scientist, August 29, 2015

Kotu, V., Deshpande, B., 2014. Predictive Analytics and Data Mining: Concepts and Practice with RapidMiner, Morgan Kaufmann

Rick Davies and Anne-Marie O'Riordan, 2015. CSCF Portfolio Analysis: Predictive Modelling of CSCF Project Achievements: Progress to Date 5th August 2015

Rick Davies, 2015. CSCF Portfolio Analysis: Predictive Modelling of CSCF Project Achievements, Updated Results

Photo (front cover)

The Education Bridge, Children dancing at a child rights dance club in Ghana

This report was prepared by Triple Line Consulting in joint venture with Crown Agents. Any views within are not necessarily held by DFID.

UKaid from the British people

TRIPLELINE CONSULTING

CROWN AGENTS

The CSCF has been managed by Triple Line and Crown Agents from 2010-2015.

This paper tests a model for the last 60 projects funded by the CSCF.