Mining Personal Banking Data to Detect Fraud

David J. Hand[1,2]

[1] Department of Mathematics, Imperial College London
South Kensington Campus, SW7 2AZ, UK
[2] Institute for Mathematical Sciences, Imperial College London
South Kensington Campus, London, SW7 2AZ, UK, *d.j.hand@imperial.ac.uk*

Abstract. Fraud detection in the retail banking sector poses some novel and challenging statistical problems. For example, the data sets are large, and yet each transaction must be examined and decisions must be made in real time, the transactions are often heterogeneous, differing substantially even within an individual account, and the data sets are typically very unbalanced, with only a tiny proportion of transactions belonging to the fraud class. We review the problem, its magnitude, and the various kinds of statistical tools have been developed for this application. The area is particularly unusual because the patterns to be detected change in response to the detection strategies which are developed: the very success of the statistical models leads to the need for new ones to be developed.

1 Background

The aim of this article is to review the application of statistical modelling ideas in the detection of fraud in the personal banking sector. The area poses some novel statistical challenges.

The *Concise Oxford Dictionary* defines fraud as 'criminal deception; the use of false representations to gain an unjust advantage. 'As such, fraud must be as old as humanity itself. Indeed, one might go so far as to claim that it is older, since even animals are known to behave in ways which deceive others, although the notion of 'criminal 'behaviour is uniquely human.

Banking fraud, in particular, has many faces. At one extreme, there is money laundering, in which one tries to pass off illegally gained funds and feed them into the legitimate banking system. At an intermediate level, there is fraud against organisations, such as commercial or public organisations. And at the far extreme there is fraud against an individual, such as through stolen or cloned credit cards. Banking fraud also covers a vast range of sizes, ranging from giant cases such as Enron and European Union fraud, to small personal cases such as selling forged tickets to soccer matches. No day passes without the national press mentioning cases of fraud - and, indeed, without countless frauds being perpetrated throughout the world.

This paper is chiefly concerned with banking fraud, and in particular fraud in the retail or personal banking sector. This covers credit cards, private

residential mortgages, car finance, personal loans, current bank accounts, savings bank accounts, and so on. It is a natural application domain for statistics and related areas of data analysis, since it involves large numbers of individual units - people.

The personal banking sector has witnessed something of a revolution in recent decades. Instead of loans and other banking products being granted by the decisions of individual bank managers, there has been a shift towards the use of objective statistical models. Such models have many advantages over humans: they do not tire or suffer from irrational changes of mood, their performance can be monitored and improved in an evolutionary way by comparing the performance of slightly modified versions, they are very quick so that one does not have to wait for days for a decision, and above all, they are consistent and no subjective or illegal prejudices can accidentally creep in. These changes have been paralleled by other changes: nowadays huge databases summarising the transaction, purchasing, and payment history of individuals is stored in computer databases. Such data warehouses provide sources of information which can be mined to better understand how people behave, and to predict how they are likely to behave in the future. And systems to obtain credit, in particular, have changed completely. In the US at the end of 2005, outstanding consumer credit, *excluding mortgages*, exceeded two trillion dollars. This is in large part the result of technical innovation. As Alan Greenspan put it in Greenspan (2005): 'Unquestionably, innovation and deregulation have vastly expanded credit availability to virtually all income classes. '

2 Personal banking fraud

With such large sums of money involved, it would be surprising if fraudsters were not attracted. The scale of the problem is illustrated by the 2005 UK figures for plastic card fraud (one can find corresponding figures for any country). The largest category of fraud was 'cardholder not present 'fraud, amounting to £183 million. This category includes phone, internet, and email fraud. The next largest was counterfeit fraud, amounting to £97 million. This includes skimming and cloning of cards, in which the electronic details are read and duplicated on another card. Close behind this was stolen or lost cards (£89 million), and this was followed by mail interception (£40 million), card identity theft from account takeover (£18 million), and card identity theft from fraudulent applications (£12 million). Of particular interest is that only the first of these, cardholder not present fraud, shows an increase over the 2004 figure. All the others show a decrease. This illustrates a particularly important point, to which I shall return below.

The figures above might be regarded as the tip of the iceberg. They represent clear direct fraud. In fact, the total loss due to fraud is much larger

because of the additional indirect components. Overall, plastic card fraud can be regarded as being composed of several components:

1. immediate direct loss due to fraud - the figures given above;
2. cost of installing and running fraud prevention and detection systems;
3. cost of loss business, for example, while a stolen card is replaced;
4. the opportunity cost of fraud prevention and detection - the other, alternative, profitable things which the money might have been spent on;
5. the deterrent effect of public fraud on the spread of e-commerce.

Little wonder, then, that some estimates give total worldwide plastic card fraud in the many billions of dollars.

3 An arms race

I commented above that all types of plastic card fraud apart from cardholder not present had shown a decrease between 2004 and 2005. This is an important point, and one which characterises statistical research in this area and introduces novel challenges. When one develops a statistical model to understand nature - in physics or biology, for example - discoveries remain true, unless or until they are replaced by more elaborate descriptions of nature which explain the data in a superior way. In fraud detection, however, this is not the case. Fraud detection represents an ongoing arms race between the fraudsters and those tasked with detecting and preventing fraud, so that the problem is inherently non-stationary. Once systems are in place to prevent a particular type of fraud, the perpetrators do not abandon their lives of crime, but move onto some other approach. We have recently witnessed a nice example of this with chip and PIN technology in the UK. Chip and PIN technology replaces signatures and magnetic stripes on cards with Personal Identification Numbers and microchips on the cards. This system was launched in the UK on 14th February 2006. Some predicted that it would reduce credit card fraud by 90%. As a consequence, it was also predicted that it would lead to an increase in identity theft (in which full financial and personal details of the victim are stolen, so that loans and other financial products, including credit cards, can be taken out without the victim being aware of it) and in fraudulent credit card use in Europe, which still relied on the signature and magnetic stripe technology. And these predictions came true - Lloyds TSB, for example, observed an increased fraudulent use of UK credit cards abroad. There was also an increase in ATM theft and cardholder not present fraud. Worse than this, however, crooks also reverted to a new use of an old technology. They had long installed 'skimmers 'in ATM machines, to record both the card details and the PIN numbers, and now they installed these in the machines used in chip and PIN systems. Over £1 million was stolen from Shell service stations before this scam was stopped.

Sleeper fraud provides another nice illustration of nonstationarity. In this scheme, fraudsters use the card in an apparently perfectly legitimate way, making transactions and repayments as if they were law-abiding users. Gradually, they ramp up their credit limit - until suddenly spending up to the limit and disappearing. It takes patience, of course, but can be lucrative, and it is very difficult to prevent.

At the time of writing, one of the newest technologies to be introduced in this arms war is the *one-time password*. There are several variants of this, but each involves using a unique password, different each time a transaction is made. This can be by using an algorithm which calculates the new password from the last one, or via time synchronised algorithms in the card and the authentication server, or in other ways. But how long will it be before fraudsters find a way round this?

4 Other challenges

If the plastic card fraud detection problem is unusual in that the characteristics of the fraud class of objects changes in response to the detection algorithms being installed, then it is also challenging in several other ways.

Generally, plastic card transaction data sets are large, often very large. If a bank has 10 million customers, making an average of 3 credit card transactions a week, then a year 's worth of transactions represents a lot of data. When one then recognises that between 70 and 80 items of information are recorded for each transaction (transaction type, date and time of transaction, amount, currency, local currency amount, merchant category, card issuer, ATM ID, POS type, number of times chip has been accessed, merchant city name, etc.) then it is easy to see that scalable and highly efficient algorithms are needed. In particular, unlike in statistical modelling, in which the aim is to produce a summary of the data which captures its distributional characteristics, so that one can use a sample of data, here it is absolutely necessary to examine each and every transaction. Dynamic updating to capture the intrinsic non-stationarity is a nice idea, but dynamic updating of millions of separate models, one for each account, as each new transaction is made, is likely to be impossible for advanced models such as support vector machines, random forests, or neural networks. Multilevel models may provide a partial answer here, in which the basic model form is the same for each customer, with just a few (easily updated) parameters being varied.

Raw fraud data sets are also typically unbalanced, having many more legitimate than fraudulent cases: an oft-quoted figure is that about 1 in a 1000 are fraudulent. This is crucial because of the familiar phenomenon, illustrated below, that high sensitivity and specificity in such cases do not translate into a high proportion of fraud cases amongst those predicted as fraudulent. The implication is that the two types of misclassification should be weighted very

differently. Multi-stage procedures can be effective approaches in such cases, as outlined below.

There is also often a delay in learning the true class labels. In fact, this is a familiar problem in the banking sector, where these labels often do not become apparent until a later reconciliation or account checking stage. It can mean a lag in updating of distributions. It is compounded with the problem of incorrect labels. There is the obvious problems that account holders may not check their statements very rigorously, so that fraudulent transactions are mislabelled as legitimate. This is a one-way misclassification, and so may not be too serious in terms of classification accuracy (its primary impact being on the classification threshold). However, consider the case of an account holder making a series of legitimate transactions, and then deciding to get the cost reimbursed by claiming that the card had been stolen and the transactions were not theirs. Now the true labels become 'fraud ', even though the transaction pattern may be indistinguishable from a legitimate series of transactions. (Fortunately, in fact, such a series would typically be distinguishable, since normally the account holder sets out to maximise their gains, and so behaves differently from normal.)

5 Statistical tools

Various statistical approaches have been explored in the battle against fraud (Bolton and Hand, 2002). Here I am using 'statistics 'in the sense of Chamberss 'greater statistics '(see Chambers (1993) and the rejoinder to Bolton and Hand (2002)), to mean 'everything related to learning from data ', so that it includes machine learning, data mining, pattern recognition, and so on. Provost (2002) makes a nice analogy with the classic parable of the blind men and the elephant - each felt a different part of the creature and imagined an entirely different sort of animal. So it is with fraud detection: there are many different approaches. It is important to recognise that these approaches are not in competition. They can (subject to scalability and computational issues) be used simultaneously and in parallel. By this means, old weapons in the fraudsters armoury will be defeated even while new ones are being tackled.

The core approach is a rule-based or pattern matching approach. This is applied when a particular type of transaction or transaction pattern is known often to be indicative of fraud. For example, the pattern of two ATM withdrawal attempts in which the first takes out the maximum allowed and the second occurs within 24 hours is suspicious. It suggests that the second attempt was not aware of the first - and that two people are using the account. Another such intrinsically suspicious pattern is the credit card purchase of many small electrical items in quick succession, since these can easily be sold on the black market. We see from these examples that one cannot be certain, merely from the transaction pattern, that fraud has occurred. A human has

to be in the loop. We shall return to this point when we consider measures for assessing the performance of fraud detection systems.

More generally, however, we will want to detect departures from normal behaviour for an individual, in unpredictable ways, as well as in predictably suspicious ways. This requires decisions about two aspects: what exactly is the unit of analysis, and what is the 'norm 'relative to which behaviour is classified as 'suspicious '?

Superficially, the unit of analysis is simple enough: it is the transaction (lying in a space with 70-80 dimensions). Sometimes people use their cards in highly predictable ways (e.g. practicing 'jamjarring ', in which they use different cards for different categories of purchase), but in other cases the transactions are highly heterogeneous. Especially in the latter case, it can be advantageous to work with groups of transactions, rather than individual transactions. This can be done in various ways. We can, for example, summarise the transactions within a group (e.g. the last 5 transactions). This allows more flexibility of description and has the potential to capture more unusual patterns of behaviour. Of course, it sacrifices the immediacy of individual transaction analysis. It also requires tools for rapid updating of the summary descriptors.

Similarly, at a superficial level, the choice of norm is straightforward: we should compare the new behaviour of a customer with his or her previous behaviour. This requires sufficient data being available on that customer previously. It also enters the realm of scalability issues: if an entirely different model has to be built for each customer then updating may be expensive. A compromise may be the multilevel approach mentioned above.

A rather different approach is to compare the behaviour of a customer with that of other similar customers. In 'peer group analysis '(Bolton and Hand (2001), Ferdousi and Maeda (2006)), for example, we identify the k customers who have behaved most similarly to a target customer in the past, and then follow them to see if the behaviour of the target customers starts to deviate from their 'peer group '. In its simplest form, this is done for each customer separately.

In the above, modelling occurs at the level of the model of behaviour which we expect a legitimate account to follow, but there is no deeper conceptualisation possible. This is to regard the account as undergoing a state change when a fraudster hijacks it, from the legitimate to the fraud state. In the former, all transactions are taken to be legitimate, but in the latter there will be fraudulent transactions, perhaps with some legitimate ones mixed in. We can think of this as a latent variable model, this variable being the state, and our aim is to detect when the state change occurs: it is a change point problem. Such problems have been extensively explored, though most often in situations in which a single manifest variable is undergoing a level shift.

Various kinds of multilevel models are particularly valuable in fraud detection problems. A straightforward application of such models is multilevel

screening. This can also help with the computation and scalability issues. In this approach, one applies a simple and quick method to eliminate the clearly non-fraudulent transactions: one computes a simple suspicion score and eliminates those with low values. Some frauds may get through, but one has to recognise that perfection is not achievable and if this initial screen can adjust the prior size of the fraud class from 0.001 to 0.01 or better then significant progress has been made. The second level may then use the same descriptive characteristics, combined in a much more elaborate and sophisticated way (e.g. using a random forest, treenet, support vector machine, or neural network) or may use additional data. The idea is analogous to the reject option, although it is one-sided.

Stolfo et al. (1997a, b) described an alternative use of multiple models, in which different fraud detection algorithms are used for different sectors, with the results being combined. Given that certain areas are more subject to fraud than others, this seems like a very sensible approach - why should one believe that the same sort of detection algorithm should apply in each area?

The aim is always to classify transactions or more general transaction groups into one of two classes: fraudulent or legitimate. Systems to achieve this can be based on supervised classification ideas, in which one uses samples of known frauds and known non-frauds to construct a rule which will allow one to assign new cases to a class (by comparing an estimated suspicion score with a threshold). But an alternative would be to estimate contours of the non-fraud class, classifying outlying points as potentially fraudulent. The contours here will most probably best be based on an individuals previous legitimate transactions. Breiman (2002) argues that the supervised approach is likely to be more effective.

So far, all of the discussion has been in terms of individual transactions, or groups of transactions within a given account, treating the accounts as independent. However, while accounts may indeed generally be independent, the ways fraudsters use accounts are not. Firstly, fraudsters tend to work in gangs, not individually (for example, stealing, recycling, and cloning multiple cards). And secondly, if a fraudster discovers a successful modus operandi, then they are likely to repeatedly use that until stopped. This can be made use of in detection systems. For example, if an account is known to have switched to the fraud state (that is, some of the transactions on an account are known to be fraudulent), one can look back at all of that accounts recent transactions and examine other accounts which made transactions at the same sites more carefully. Quite how effective this will be will depend on what data are stored about the transactions. If individual ATM identifiers are stored, it will be easy for ATM transactions, for example. If only high level merchant codes are stored for credit card transactions, however, then it would result in a much blunter instrument. The idea is a dynamic version

of simple methods based on learning what merchant codes are intrinsically more likely to be associated with fraud.

6 Assessing performance of fraud detection tools

Although different kinds of techniques may be used to process a transaction or activity record, the aim in all cases is to assign them to one of two classes, fraud or non-fraud. This is even the case if the problem is viewed as one of detecting state change: one aims to classify those prior to the change as legitimate and those after the change as fraudulent. This means that an important class of performance assessment measures must be based on the two by two cross-classification of true class (fraud, non-fraud) by predicted class.

The classification community has developed many measures for such situations, tackling different aspects of performance. Simple ones include misclassification rate and specificity and sensitivity. As we have already mentioned, these are typically inappropriate in fraud detection problems because of the dramatically unbalanced class sizes: a very low misclassification rate (0.1% if only 0.1% of the transactions really are fraudulent) is achieved by assigning every transaction to the legitimate class. But this, of course, defeats the object. The point is that misclassifying a fraud case is much more serious than misclassifying a legitimate case. The former means a real financial loss, which could run into many thousands of pounds. The latter incurs only the cost of checking that the transaction is legitimate, plus also some customer irritation if the account is temporarily suspended. This irritation can be managed - after all, most customers like to know that the bank is looking out for them. If the true fraud rate is 0.1% then a detection rule which successfully classifies 99% of the fraud cases as fraudulent, and 99% of the legitimate cases as legitimate will in fact be correct in only 9% of the cases it predicts as fraudulent. This could mean substantial customer irritation, not to mention the cost 'wasted 'on the 91 in every 100 suspected frauds which are really legitimate.

There are also other aspects of fraud performance which one might want to take into account. Hand et al. (2006) point out that whenever a fraud is suspected, it incurs an investigation cost, regardless of whether a fraud has actually been committed or not. Thus a suitable measure might be based on minimising a suitably weighted combination of the total number of fraud alarms and the number of real frauds which evade detection. Even more elaborate measures may be based on the actual monetary losses incurred when a fraud does occur.

7 Conclusion

At a conference on banking fraud I attended not so long ago, a banker remarked to me that his bank 'did not have any fraud '. He was speaking tongue in cheek, of course, but some important points underlie his comment.

The first is that it is very important, for customer and shareholder confidence, to know that a bank is a reliable organisation, not subject to criminal attacks, and to the costs that that would imply. The contrary assertion (or, perhaps, admission) that the bank loses hundreds of millions of dollars per annum to fraud would hardly inspire confidence.

Secondly, at a superficial level there would appear to be an appropriate balance to be struck between the amount spent on detecting and preventing fraud and the amount of fraud prevented. One might decide that a break-even point was appropriate: it might be regarded as sensible to spend £x to prevent £x of fraud, but foolish to spend £y to prevent £x if $y > x$. This is all very well, but it ignores the deterrent effect: a fraud system costing £y may prevent substantially larger amounts of fraud merely because it is known to exist - merely because the bank is known to be able to detect fraud attacks.

In any case, while one might be able to quantify the amount spent on fraud detection and prevention systems, quantifying the amount saved by these systems is difficult. After all, if fraud is not attempted by virtue of a prevention strategy, how can its extent be measured? In general, quantifying the value of fraud detection systems is difficult.

I commented above that once a particular avenue of fraud has been prevented by an appropriate tool, fraudsters do not abandon their efforts, but change tacks. This means that a Pareto principle applies. 50% of fraud is easy to detect - the early methods used by those new to the game. But the next 25% is much harder, and the next 12% harder still. Indeed, it would be naive to suppose that all fraud is prevented or could be prevented, no matter how sophisticated the statistical models. Think of those previously law-abiding customers who suddenly realise that, if they claim their card has been stolen after a spending spree, they will be reimbursed. Think of sleeper fraud.

Other reviews of statistical approaches to fraud detection are given in Fawcett and Provost (2002), Bolton and Hand (2002) and Phua et al. (2005).

Acknowledgments

This work has been partially supported by EPSRC grant number EP/C532589/1 for the *Thinkcrime* project on *Statistical and machine learning tools for plastic card and other personal fraud detection.*

References

BOLTON, R.J. and HAND, D.J. (2001): Peer group analysis. Technical Report, Department of Mathematics, Imperial College, London.

BOLTON, R.J. and HAND, D.J. (2002): Statistical fraud detection: a review. *Statistical Science, 17, 235-255.*

BREIMAN, L. (2002): Comment on Bolton and Hand (2002). *Statistical Science, 17, 252-254.*

CHAMBERS, J.M. (1993): Greater or lesser statistics: a choice for future research. *Statistics and Computing, 3, 182-184.*

FAWCETT, T. and PROVOST, F. (2002): Fraud detection. In: W. Kloesgen and J. Zytkow (Eds.): *Handbook of Knowledge Discovery and Data Mining*, Oxford University Press, Oxford.

FERDOUSI, Z. and MAEDA, A (2006): Unsupervised outlier detection in time series data. In: *Proceedings of the 22nd International Conference on Data Engineering Workshops*, ICDEW06, IEEE, 51–56.

GREENSPAN, A. (2005): *Consumer finance.* Remarks presented at the Federal Reserve Systems Fourth Annual Community Affairs Research Conference, Washington DC, 8th April.

HAND, D.J., WHITROW, C., ADAMS, N.M., JUSZCZAK, P., and WESTON, D. (2006): Performance criteria for plastic card fraud detection tools. To appear in *Journal of the Operational Research Society.*

PHUA, C., LEE, V., SMITH, K, and GAYLER, R. (2005): A comprehensive survey of data mining-based fraud detection research. Technical Report, Monash University.

PROVOST, F. (2002): Comment on Statistical fraud detection: a review. *Statistical Science, 17, 249-251.*

STOLFO, S., FAN, W., LEE, W., PRODROMIDIS, A.L. and CHAN, P. (1997a): Credit card fraud detection using meta-learning: issues and initial results. In: *AAAI Workshop on AI Approaches to Fraud Detection and Risk Management*, AAAI Press, Menlo Park, CA, 83–90.

STOLFO, S.J., PRODROMIDIS, A. L., TSELEPIS, S., LEE, W., FAN, D.W., and CHANN, P.K. (1997b): JAM: Java agents for meta-learning over distributed databases. In: *AAAI Workshop on AI approaches to Fraud Detection and Risk Management*, AAAI Press, Menlo Park, CA, 91–98.