Use of Kohonen Self-Organizing Maps and Behavioral Analytics to identify cross - border smuggling activity

Grant M Brown

***Abstract*—Risk assessment of movements entering the UK using Kohonen SOMs and decision trees was found to be 292% more accurate at discerning smuggling than incumbent systems. Inbound freight is currently assessed using a series of risk flags based upon data provided by ferry companies to Government border agencies and a previous offender watchlist. This data constituted the input for the analysis, but presented complex challenges; including proportionally few seizures to train predictive models and a relatively un-diverse "selected for search" population. These factors prompted use of an unsupervised clustering technique to uncover abnormalities in the data and quantify changes in behavior. A supervised technique (a binary decision tree) was also used to enhance existing profiles. Initial tests indicated movements resulting in seizures had a Euclidean distance from their average centroid 10 times greater than the non-seizure mean and only 5% of vehicles ever moved significantly from their average centroid. Subsequent blind tests "selected" the riskiest 20% of movements - from which 39% of all seizures and 57% of high value seizures occurring over the same period were identified. Based upon this research, further work has been commissioned to design a system which uses these methods for targeting other movement types.**

***Index Terms*— Kohonen Self-Organizing Maps, decision trees, unsupervised analysis, government applications, SAS.**

I. Introduction

Cigarette and drug smuggling are issues of great concern for the United Kingdom Government: cigarette smuggling in particular accounts for an estimated £2.5 Billion a year in lost tax revenue. This allows smugglers to make large profits on cheap imported cigarettes to the detriment of the UK government [1].

One of the easiest ways to transport goods into the UK is via lorry and Roll-On/Roll-Off (RoRo) cross channel ferries and the Channel Tunnel. The Government department who polices these borders has responsibility for prosecuting the drivers of these vehicles as well as detecting contraband. Increasing seizure yields while at the same time facilitating the free movement of legitimate traders has forced the Client to review more advanced techniques for identifying illegitimate activity.

A. *The problem*

Historically vehicles were selected for search based on intelligence profiles. In 2005 these profiles were automated by a Freight Targeting System (FTS) which also utilizes vehicle licensing and UK taxation data to aid targeting. While this has improved targeting there are still problems with this approach:

- Current data analysis is of the seizure population only, which can lead to self-fulfilling prophecies. For example, if 70% of seizures come from red vehicles, it could be deduced that 'red vehicle' is a good indicator of criminality. If 70% of selected vehicles are red however, we can see that this indicator has no discriminatory power;
- Many of the risk indicators can only be applied to UK vehicles - which form ~15% of the inbound traffic - as they are based on haulage company tax information and vehicle licensing agency data;
- Difference between non-compliance and legitimacy is a continuum. The current profile indicators are binary in their ability to separate populations of risk. For example, early booking might be an indicator of risk; but the amount of risk varies depending on how close to the sailing a booking takes place;
- Consistent risk scoring methodology is required. Risk scoring is currently not trusted; statistically accurate risk scoring would considerably improve targeting consistency and effectiveness;
- No succinct way exists to compare how a movement differs from the norm; by segmenting vehicles we can assess risk by scoring movements to different segments and flagging vehicles which appear in high risk segments;
- There are patterns of risk which FTS and existing systems cannot detect.

The greatest advantage with the implementation of FTS it that is has allowed a vast amount of data to be collected – not just on vehicles associated with seizures (as in the past) – but across all movements in and out of the UK. This has allowed the use of more advanced analysis techniques.

Manuscript accepted April 24, 2007. This work was commissioned by a United Kingdom Government agency as part of the Freight Targeting System project and the wider eFrontiers initiative. The work was contracted to Detica PLC.

Grant Brown is a senior consultant with Detica PLC, 2 Arundel Street, London, WC2R 3AZ, United Kingdom, (Tel: +44 7979 521 936; e-mail: grant.brown@detica.com).

B. *Techniques used*

Analysis had shown the ‘selected vehicle’ population was relatively un-diverse and thus an unsupervised technique was chosen. Customs Officers asserted that generally there was a change in behavior associated with vehicles prior to a seizure. This made it clear that a clustering method which aggregated a large number of metrics, and generated a measurable distance between clusters would be required. Whilst there are many clustering techniques, the Self-Organizing Maps (SOMs) were seen as preferential because like all neural network techniques the most important input vectors in deciding the formation of clusters are magnified through the learning process. As the data was of low quality and the underlying defined behaviors only had limited discriminatory power individually, a technique which ‘learned’ the most important combinations of behaviors with regards to differentiating groups was seen as preferential. Additionally the topographical nature of Kohonen SOM’s made the distance scores that would be outputted by a vehicle changing clusters more meaningful.

It was also felt that a supervised technique could be used to show what traits in the vehicles currently selected were statistically significant within the seizure population and therefore good indicators. The data was thus classified using a binary decision tree. The techniques were then trialed between the 1-14 October 2006 and achieved a selection to seizure conversion rate 292% better then existing methods.

There were other indicators of criminality, which were analyzed using Social Network Analysis. This work is detailed in a separate paper.

II. METHOD

The first step was to define the behaviors which could be extracted from the available data sources. The most consistent data available was Ferry Company ticketing information and this formed the core of our behaviors, which were thus heavily based upon Customer Relationship Management (CRM) practice. The primary behaviors fell into the following categories:

- Recency (e.g. Days since last movement, days since last arrival)
- Latency (e.g. Average days between movements, standard deviation of time between booking tickets and sailing)
- Frequency (e.g. Number of appearances at this port, % of occasions this vehicle carries foodstuffs)
- Quantity (Trend information; e.g. how common is it for a vehicle to appear from this ferry)

In addition there are a number of second order behavioral metrics which compare a given vehicles’ first order metrics for one movement with the same vehicles’ average metrics.

We also wished to compare a vehicle with other vehicles at a haulage company, but the poor data quality prevented this. The vehicle was compared with other vehicles using the same booking account however, and whilst not perfect this did have some discriminatory value.

Most of the behaviors are centered on the vehicle as the registration data was a strong key to match with previous movements. Accurate matching of individuals with historical data was difficult as only ~30% of ferry operators provide driver passport numbers. Names alone were not a strong enough key for accurate matching – even with the use of SSA[1] ‘fuzzy’ matching technology.

The behaviors were built from the data using SAS linked to an Oracle database. In total there were 60 metrics created. This allowed analysis of the behaviors predictive/discriminatory power. This in itself was a useful exercise as many of the indicators previously relied upon were proven to have little discriminatory power.

From these behaviors we built a decision tree and Kohonen Self Organizing Map. SAS Enterprise Miner 5.2 has dedicated nodes that allow for the creation of these by the user. The methodology followed SAS’s SEMMA (Sample Explore, Modify, Model and Asses) guidelines. One of the many advantages of using this tool is that the effectiveness of various statistical techniques, such as correspondence analysis and factor analysis could be assessed without significant effort.

A. *Kohonen Self Organizing Map (SOM)*

SOM’s [2,3] provide a way of representing multidimensional data in much lower dimensional spaces. They are ideally suited to exploratory data analysis, allowing one to impose partial structure on the clusters as well as better facilitate visualization and interpretation for users. Mangiameli et al [4] applied SOMs and seven hierarchical methods to 252 ‘‘messy’’ data sets with real-world data imperfections. SOMs were found to be superior in both robustness and accuracy.

To create an SOM, one first chooses a geometry of ‘‘nodes’’— here a 5 by 4 grid, leading to 20 clusters, was used. The nodes are mapped into k-dimensional space, initially at random, and then iteratively adjusted. Each iteration involves randomly selecting a data point P and moving the nodes in the direction of P. The closest node NP is moved the most, whereas other nodes are moved by smaller amounts depending on their distance from NP in the initial geometry. In this fashion, neighboring points in the initial geometry tend to be mapped to nearby points in k-dimensional space. The process continues for 20,000–50,000 iterations. Any new, previously unseen input vectors presented to the network will stimulate nodes in the zone with similar weight vectors.

The mapping of nodes is adjusted by moving points toward P by the formula:

$$f_{i+1}(N) = f_i(N) + \tau\,(d(N, N_p), i)(P - f_i(N))$$

The learning rate τ decreases with the distance of node N from N_P and with iteration number *i*. The point P used at each iteration is determined by random ordering of the *n* data points generated once and recycled as needed. The function τ is

[1] SSA-NAME3 is a Commercial of the Shelf (COTS) product which can be embedded into J2EE and Oracle systems and allows ‘fuzzy’ matching, effectively dealing with textual errors. For more details see www.identitysystems.com/products/name3T.htm

defined by

$$\tau (x,i) = 0.02T/(T/100\ i)$$

for $x = \rho(i)$ and $\tau (x, i) = 0$ otherwise, where radius $\rho(i)$ decreases linearly with $i(\rho(0) = 3)$ and eventually becomes zero [5]. Once it becomes zero the maximum number of iterations have been completed and the data clustered; T is the maximum number of iterations.

As stated, the data was classified into 20 clusters, which mapped to different traveller types. Each cluster contained a roughly equal proportion of the whole population, ranging between 2 and 15%. A grid was developed to show the users what kind of traveller occupied a certain cluster, which was found to be very useful. For example cluster 20 was "*Uses very large account. Above average proportion of loaded movements. Above average number of movements for this vehicle, and frequency pattern correlates well with the other vehicles which use this account.*"

Figure 1 shows how movements can be risk scored by their Euclidian distance from their normal behavior segment, and the distance of the clusters generated in this exercise. It should be noted that few movements would be found in the white space outside of the clusters.

B. *Decision Tree*

In a decision tree, each stage in the tree is a question asking whether a given movement has risky attributes. It terminates at a leaf when a decision about the movements risk can be discerned.

The selected population was broken into three groups for this analysis; 0 was no seizure, 1 a low value seizure and 2 was a high value seizure (class A narcotics or over 100,000 cigarettes). Each leaf has a percentage attributed to each of these three groups. The tree was 'trained' using a profit mechanism and a 3 month slice of movement data. If the tree predicted a 2 (a high value seizure) and a 2 was the result, a profit of 1200 was awarded. Conversely if the tree predicted a large seizure and no seizure resulted, a fine of -700 was given. Using this method we could calculate the most profitable combination(s) of behaviors. Whilst a certain combination of rules may have predicted vast numbers of seizures, it may also have generated an even larger number of false positives, resulting in a lower profit.

Profit scores were used rather then probabilities because of the small size of the seizure population. If we had used probabilities, the majority of leaves would have predicted that no seizure would occur. The other way around this problem was to boost the seizure sample relative to the whole population. When we used this method however, the models built tended to be over-fit. The profit awarded for each outcome was arrived at via trial and error; multiple iterations helping to create discriminatory trees which were robust and not over-fit.

The decision tree builds itself from the data. The risk metric used at each branch is the one with the most discriminatory power at that point, based upon the attributes of the remaining population. Because we used a binary decision tree, the same metric could be used multiple times, each step through the tree filtering the population further. The tree was built using data from November 2005 to May 15th 2006 – data from May 15th to June 6th 2006 was set aside as a test data set.

C. *Testing the models*

The first phase of testing these techniques considered both separately. The tests used the historical data captured between May 15th and June 6th 2006. First, we compared seizure results and selection information with the decision tree's predictions over the same three week period. The SOM, which was designed to show abnormality rather then predict seizures, was more difficult to test. Using the historical data, we were able to show that very few vehicles changed behavior between movements. Additionally the SOM identified unusual changes in behavior in a number of case studies from recent large seizures.

The decision tree and Kohonen SOM were introduced to the client on the 25th September 2006, where our positive results from phase one led to the client issuing a challenge. With only movement data, and no seizure or selection data, we had to predict the seizures which took place between the 1st and 14th October 2006. We risk scored all movements using both

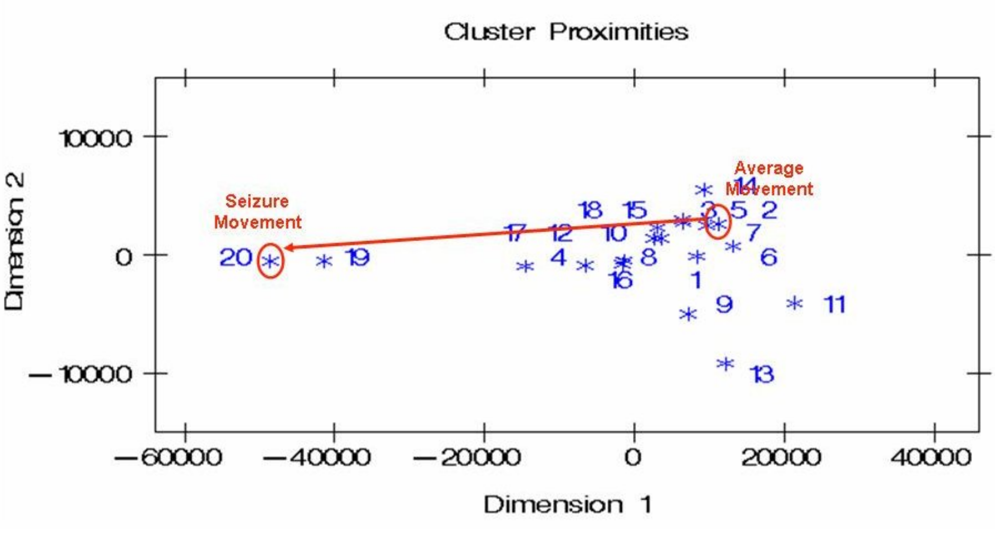

Fig.1 Using Kohonen clusters to generate risk scores

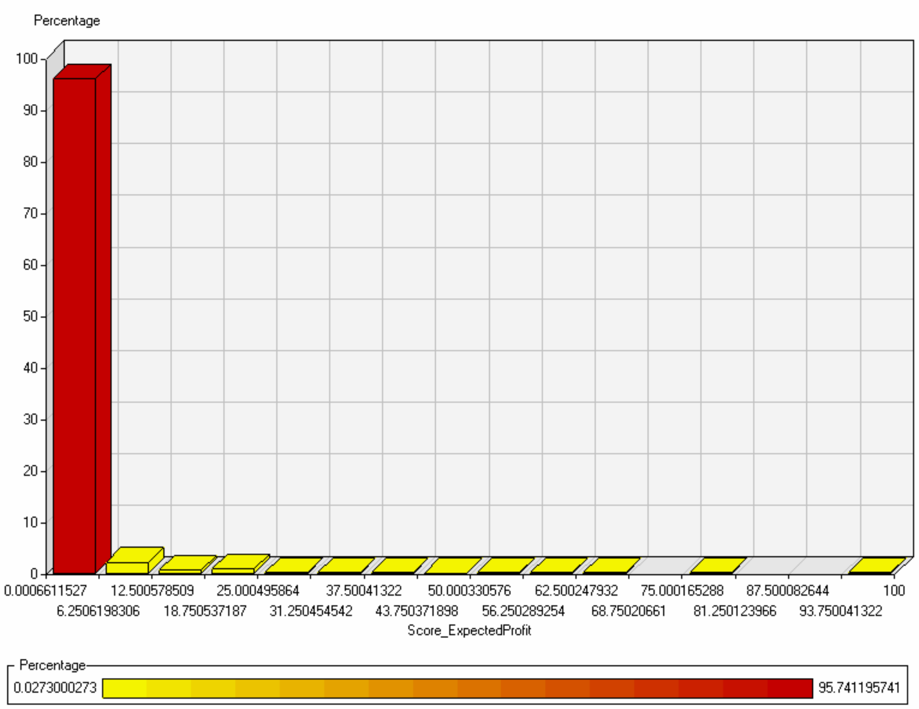

Fig. 2. Decision tree profit score

methods during this period. The risk scored movement data was loaded into an Access database and a weighted final score generated; we defined the top 20% as worthy of selection. While this was a larger percentage then was selected by the Dover teams, if our selections were as good as random we would only be expected to uncover 20% of the seizures.

Targeting effectiveness could be discerned by dividing the percentage of seizures by 20%. Random would equal 100, greater then 100 would show the techniques had some discriminatory power.

III. RESULTS

The results, like the tests, are presented in two phases; a description of the initial test results from the two methods using the historical data followed by the results of the Detection and Intelligence departments challenge.

The initial test of the decision tree showed that:

- 40% of all seizures would have been flagged as high risk movements by the decision tree;
- 82% of selections that didn't result in seizures would have been flagged as low risk by the decision tree;
- Only 5-6% of incoming vehicles were identified as high risk; thus the tree did not generate large numbers of false positives (see figure 2).

The Kohonen SOM clustering technique assessed risk based on whether a vehicle/driver was behaving differently to their usual modus operandi. Initial results showed that:

- Only 5% of vehicles regularly change behaviors; again showing the technique is unlikely to generate masses of false positives (see figure 3);
- When the seizure population was superimposed into the technique it was noted that vehicles involved in seizures had a Euclidean distance from their average centroid ~10 times greater then the mean for non-seizure movements distance from their average centroid

The above result means a distance score would have strong discriminatory power as a risk indicator. As stated the majority of movements very rarely stray from their cluster; 95% always remain in the same cluster, and 80% do not even move noticeably from their usual centroid. These two facts, combined with the results from the decision tree test, contributed strongly to the Detection and Intelligence Department's decision to combine these two methods into a single model and test the model in what became known as the Detection Challenge.

Using only movement data, the model produced a set of predictions. These results were compared to the actual seizures in order to establish how effective the techniques had proved. The model predicted 39% of all seizures, but more significantly the model predicted 57% of seizures above the National Intelligence debrief threshold – i.e. seizures of over 250,000 cigarettes and class A narcotics. These larger seizures were the primary focus of this evaluation since smaller seizures should be harder to predict; they are rarely done by organizations or are pre-mediated. As such the underlying behaviors are more likely to be highly variable. The effectiveness[2] of the model and that of the current selection procedure were then calculated.

[2] Effectiveness compares the percentage a trait appears in the seizure population versus how often a trait appears in the selected vehicle population. A trait which has no discriminatory value would appear equally in both populations, and would have an effectiveness of 100. 20% would be random. 38% was achieved during the trial, 38/20 x 100 = 190.

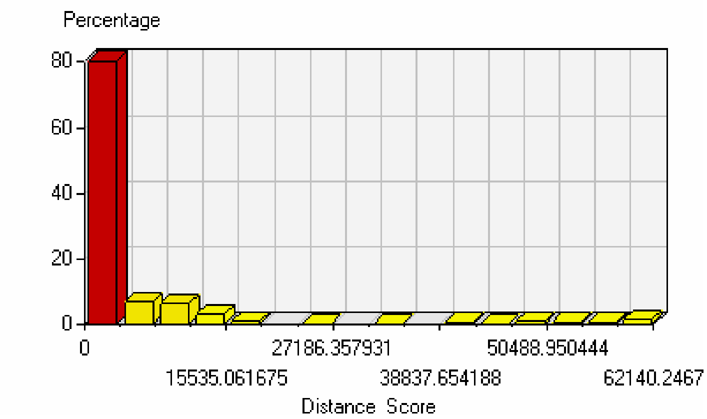

Fig. 3. Distribution of distances from average centroid

The models were found to have an effectiveness of 190, compared with the existing FTS pilots' effectiveness of 112 – a 71% improvement. This is shown in Figure 4. The red line indicates the conversion rate[3] associated with a level of effectiveness based on current performance. This conversion rate is not an exact figure, but a guideline based upon the conversion rates associated with different FTS rules with a given level of effectiveness. It shows that the new model could bring a substantial improvement in conversion rate.

There was a significant divergence in what was deemed risky by the Detection teams and the behavioral model. For example, the model selected 13% of what Detection selected; significantly fewer than the 20% random sample. Therefore, out of all the movements which were selected, the model only judged 13% of the vehicles searched to be high risk. From this 13%, 39% of all the seizures were identified. If the behavioral techniques were as accurate as random selection, they would only have identified 13% of the seizures; by predicting 39% of the seizures, they have been shown to be 3 times more accurate then existing targeting practices, at least for targeting a high risk subset of the seizure population.

17 seizures were predicted by the model from the 395 selections that the model would have selected which were actually rummaged; a conversion rate of 4.3%. This compares favorably with Detection's conversion rate of 1.47% over the same period – thus the techniques demonstrated an improvement of 292%. Of course we cannot predict the outcome of the other selections, but this evidence was sufficient for the client to commission a system which uses these methods for targeting cross-border traffic. Even with the limitations of the current data, these techniques could help the client reach their conversion rate target of 5%. With the improved data promised by the Immigration, Asylum and Nationality (IAN) bill, these techniques could prove even more accurate.

IV. DISCUSSION

These techniques could bring great benefit to anti-smuggling operations. If the client were able to achieve a conversion rate of 4.37% operationally, with the seizures being of the same average value, the client could expect to increase the number of cigarettes seized by 302%, an increase worth £16.3 Million per annum. Even if the models were half as effective operationally as they were in the trial, the client could expect to seize 56% more cigarettes, an increase worth £4.3 million.

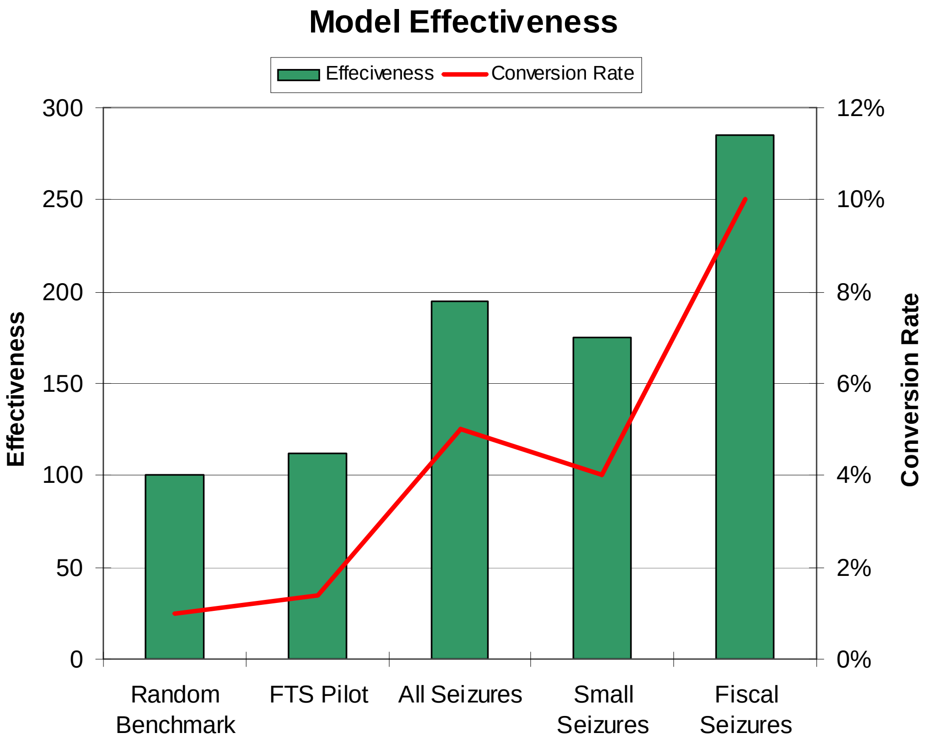

Fig. 4. Behavioral models targeting effectiveness

[3] Conversion rate is simply the percentage of vehicles stopped which lead to seizures.

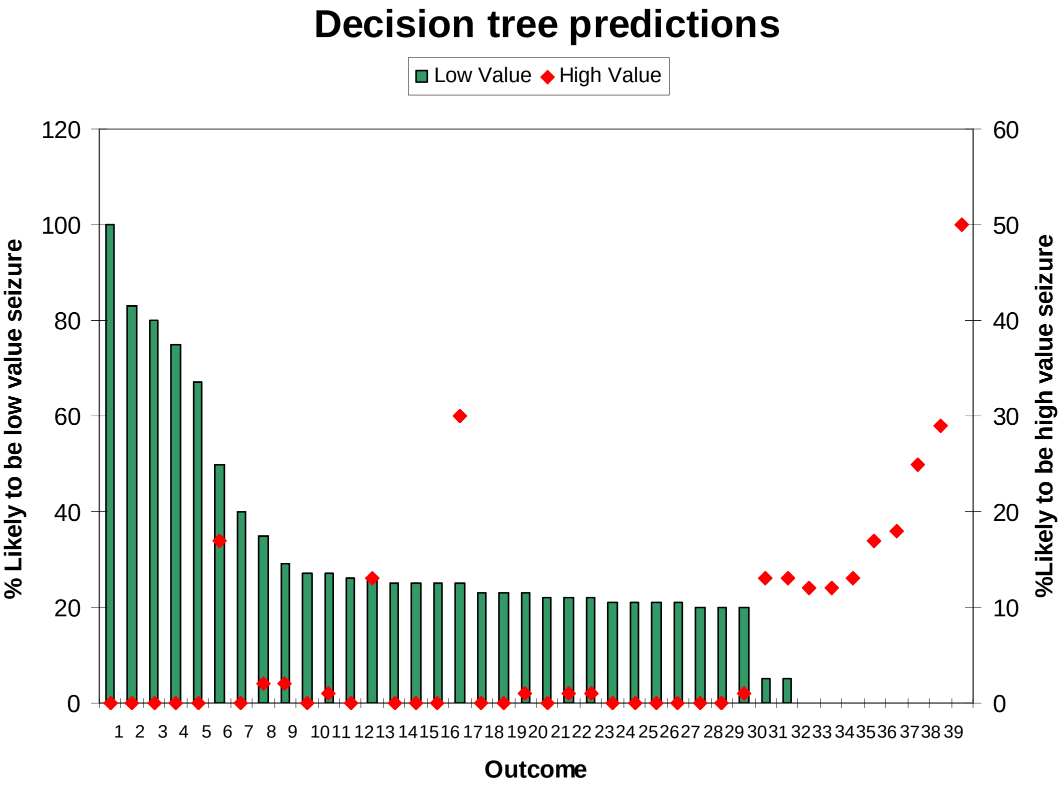

Fig. 5. Discriminating between high and low value seizures

Whilst it could be argued that the models identified those seizures which were the easiest to identify, the ability to avoid false positives and quickly identify movements which are high risk via the techniques described, means the system will free up resources for targeting the more difficult to identify smugglers.

Additionally, the models will undergo a circle of continuous improvement resulting from using these techniques and then rebuilding the models from new data with a greater number of seizures to drive predictions. These factors could realistically lead to better conversion rates. A 6% conversion rate, for example, would lead to the seizure of an additional 144 million cigarettes a year.

In addition to predicting more seizures, the models were shown to be better at finding high value seizures. The decision trees created for the Challenge exercise were specifically designed to search for seizures of over 100,000 cigarettes.

The leaves which predicted a high value seizure very rarely predicted a low value seizure. Figure 5 shows the predicted outcome of the 40 highest profit leaves in the tree. The green bars show the percentage probability that a movement in that leaf is likely to result in a small seizure if the vehicle is stopped. The red points show the probability that a movement within that leaf is likely to result in a large (>100,000 cigarettes or Class A drugs) seizure. Crucially when the tree predicted a small outcome it rarely predicted a large outcome and vice versa.

Despite only accounting for 25% of all seizures, seizures with more than 100,000 cigarettes account for 98% of all cigarettes captured. If – by actively targeting them – large seizures comprised 30% of the seizure population, the Client would increase the value of it's annual seizures by 50%.

ACKNOWLEDGMENT

Robert Hargraves for performing a large amount of the data analysis, Richard Haycock for leading the FTS project, Daniel Telehagen for extracting the data from FTS and various other systems, Bryan O'Neil for technical support and Melissa Sullivan for proofing and clarity suggestions.

REFERENCES

[1] Source: Her Majesty's Treasury 2005 budget press notice, http://www.hm-treasury.gov.uk/budget/budget_05/press_notices/bud_bud05_press03.cfm

[2] Kohonen, T.Proceedings of the IEEE 78, 1991 1464–1480.

[3] Kohonen, T. Self-Organizing Maps, Springer, Berlin, 1997

[4] Mangiameli, P., Chen, S. K. & West, D. "Ensemble strategies for a medical diagnostic decision support system," European journal of operational research .162, 2005, 432-551.

[5] Tamayo et al. "Interpreting patterns of gene expression with self-organizing maps: Methods and application to haematopoietic differentiation," Proceedings of the National Academy of Science, USA, Vol. 96, 1999, 2907–2912,