Paper 261

Use of Enterprise Miner™ & Multivariate Statistical Analysis to Build Fraud-Detection Models

Shiao-ping Lu, Lucid Consulting Group, Palo Alto, CA

ABSTRACT

The increasing number of checking account fraud cases and the dollar loss reported at Wells Fargo Bank evoked the need to build an early-detection program at the point of transaction. Since the fraud rate is relatively small compared to millions of transactions per month, it is both desirable and warranted to develop new robust models and to provide early warnings to the fraud prevention unit.

This paper will address the fraudulent behavior, basic algorithm, sampling, model building and results assessment. Predictive modeling techniques such as logistic regression, tree-analysis and neural network are illustrated with the use of SAS® Enterprise Miner™. The accuracy of prediction, misclassification costs and robustness of various models will be compared.

INTRODUCTION

Background

In spite the presence of several fraud detection programs, the reported cases of checking account fraud and millions of dollar loss are rising at Wells Fargo Bank. Majority of the fraud detection is operated in batch processing (after-the-fact) and the detection mechanism is more or less "chasing the fraud". A typical scenario is once the organized criminal group changes the modes of operation, the existing programs are no longer effective. In order to bring loss under control and to minimize it, a new strategy must be implemented. This paper will discuss the basic concepts and illustrate some of the implementation techniques by utilizing SAS Enterprise Miner™**.**

Two major types of fraud, victimized and perpetrating, have been reported based on the modes of operation. Transactions go through various channels, over the counter, transit, ATM etc. to be processed by the bank. Thus far, major dollar loss and number of fraud cases have come from over-the-counter (OTC) channel because transactions can go through in a shorter time window. This paper will illustrate the model building on the victimized frauds. A victimized fraud is defined as the crooks use several techniques such as making counterfeit check, forging signature, altering the check and dollar amount on a written check or stealing the checks and cashing out from a victimized account.

Desirable Model Performance

The ratio of business accounts Vs consumer accounts is 1 to 10 while that of business frauds to consumer frauds is 3 to 2. Preliminary analyses suggested building separate model for business and consumer accounts. Since the overall fraud rate, approximately 1 in 10,000 transactions, is relatively small compared to over 10 millions of transaction per month, the model needs to be robust enough to yield high hit rate while keeps the false positive rate low. If the latter is not low, major operation inconvenience and customer complaints will be resulted. The definitions for hit rate, false positive and false warning rate are tabulated in Table 1. Taking into consideration the number of branches and transaction volume generated per day, our goal is to keep the false positive rate around 5% and 3% for business accounts and consumer accounts respectively. On the other hand, a high false warning rate (the proportion of predicted frauds that are in fact legitimate) is acceptable because the cost of managing a false warning case is negligible compared to an average fraud loss.

Table 1
Definition of Model Performance Statistics

		Actual		Totals (Actual)
		F	NF	
Model Predicts	F	a	b	a + b
	NF	c	d	c + d
Totals (Predicts)		a + c	b + d	N

Property	Definition	Computation from Table 1
Sensitivity (Hit Rate)	Probability that an actual fraud case is predicted as a fraud	$\frac{a}{a+c}$
False Positive Rate	Probability of known legitimate transactions that are predicted as fraud	$\frac{b}{b+d}$
False Warning Rate	Probability of predicted frauds that are actual legitimate	$\frac{b}{a+b}$

Strategy for Algorithm Development

The modeling strategy is to look for the deviation of fraudulent transaction behavior from normal customers' behavior pattern. Therefore, a year's worth of transaction history for each sample account was constructed, and such deviations in transaction behavior are coded as predictors in the predictive models. We believe such models will prevent good customers from being victimized and do not depend on knowing in advance what the fraud perpetrators are attempting.

SAMPLING AND ANALYSIS FILE PREPARATION

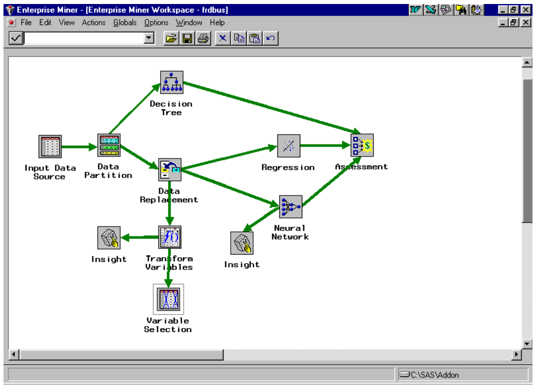
Given the aforementioned imbalance of consumer Vs business accounts, a stratified sampling approach was used to sample equal number of business and consumer accounts. For these accounts, a one-year worth of transaction history was built and several transaction variables were derived per business-cycle basis. A business- cycle is equivalent to one-month lengthwise except the cycle starting date and ending date vary from account to account. From these accounts, 4000 legitimate transactions were then randomly sampled by account ownership for modeling.

Several thousands of fraudulent transactions occurred in 1997 & 1998 were included and the transaction behavior variables were derived in a similar way as legitimate transactions for each fraud case. 4511 business frauds and 2864 consumer frauds were used for modeling.

DATA EXPLORATION AND MODIFICATION

SAS Enterprise Miner®. was employed to build the predictive models through various multivariate statistical approaches: neural networks, logistic regression and decision tree. The diagram shown in **Figure 1** illustrates an example of steps taken for business account modeling. The Input Data Source Node requires assignment of target and input variables for the model. The distribution of both continuous (interval) and categorical (class) variables can be viewed to determine the extent of missing data and transformation that might be required. For logistic regression and neural network modeling, data replacement node was required to impute two input variables with 14% and 29% missing respectively. Further data transformation was performed for two purposes: to enable distribution of the transformed variable approximate normality and to discretize the interval variables by creating "bins" or "buckets" based on quantiles before feeding the variable to the logistic regression or neural network as an input.

FIGURE 1



MODELING

Neural Network Model

An artificial neural network can be defined as a computer application that attempts to mimic the neurophysiology of the human brain in the sense that it learns from examples to find patterns in data. By finding complex non-linear relationship in data, neural networks can help to make predictions about real-world problems. The data was partitioned 50: 50 to training and validation respectively. Several hidden layers and number of iterations were experimented**.** In addition, the prior probability of fraud, 0.0001, was specified as weights in model manager to make adjustments to the posterior probabilities computation and results assessment. **Figure 2-4** presents the lift charts (a.k.a. gains chart) of neural network with 4 neurons. The horizontal axis gives the centile groupings of the predicted probability (the highest are on the left). Because of the low fraud rate, the h-axis is grouped in centile instead of decile in order to capture the fraud rate within the top 10% cases. The vertical axis of **Figure 2** represents the actual fraud rate in each centile. If the top 1% of the cases were targeted, then the fraud rate would be greater than 0.90% compared to the baseline fraud rate around 0.01% (with a lift value of greater than 90). Lift value, shown **in Figure 3** is computed as cumulative response rate divided by baseline response rate. The fluctuation represents random variation. **Figure 4** presents a Lorenz curve (also known as ROC curve) which depicts the cumulative percent of actual frauds that are captured in each cumulative centile. It shows 92% of all frauds would be captured, if the top 10**%** of the cases were targeted.

FIGURE 2

% Cumulative Response

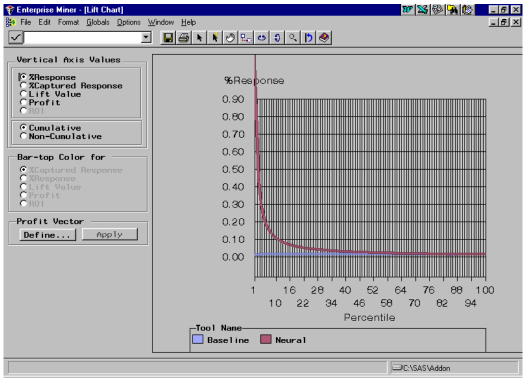


FIGURE 3

Lift Chart

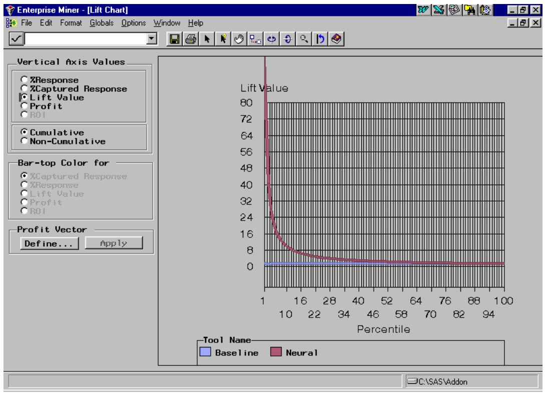
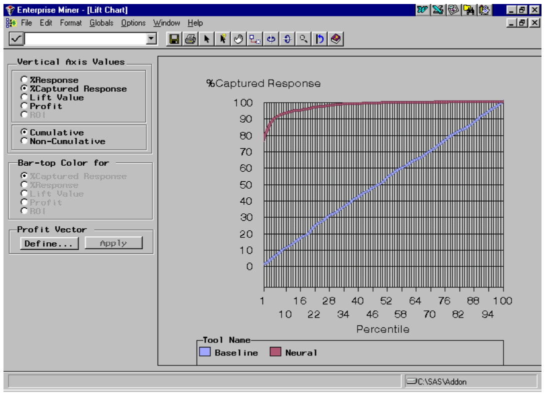


FIGURE 4

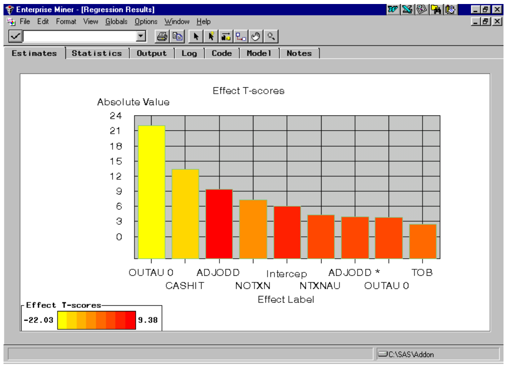
% Cumulative Captured Response



Logistic Regression

Through Variable Selection node, the R-square (measurement of model goodness-of-fit) improvements with the addition of each main effect and their two-way interactions were evaluated. All main effects plus three two-way interactions were fed into the logit model**.** **Figure 5** depicts an example of the significant parameter estimates (standardized known as effects T-score) to the response which makes it easier-to-understand than the neural net model.

FIGURE 5
Effects T-Score



Tree Based Models

Since missing values can be treated as another category of the analysis variable, no data imputation was performed. **Figure 6** shows a partial tree diagram; the "root" of the tree is the entire data set. The subsets form the "branches" of the tree. Subsets that meet a stopping criterion and thus are not partitioned are "leaves". Any subset in the tree, including the root or leaves, is a "node". The decision tree contains 18 nodes. Four subsets (or leaves) were found to contain more than 75% of the fraud; the larger leave has 2295 observations and the small one has 136 observations.

Figure 6
Partial Tree Diagram

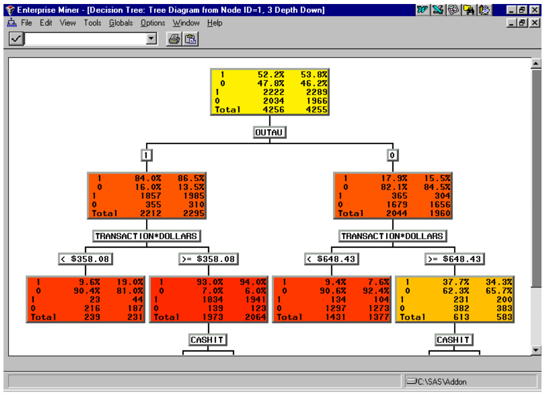


Table 2 tabulates the hit rate (sensitivity) and misclassification rate by various models for business frauds using posterior probability of 0.75 as the cutoff. Neural net model yielded a workable false positive rate of 3.15% while achieving 85% hit rate. Both tree model and logistic regression gave excellent hit rate at a slightly higher false positive rate of 5.9% and 4.7% respectively. The decision to go with which model lies in several considerations: ease of large-scale implementation, cost of missing a fraud and the tolerance towards false positive rate in the branch operation.

In the case of consumer fraud modeling, neural net model outperforms logistic regression and tree-based model in term of % response rate and % captured response rate. At 0.75 cutoff, the hit rate is 72% with 5% false positive rate as planned.

Table 2
Two-way Classification Table

Neural Network Model

Actual	Predict	
Frequency Percent Row Percent	0	1
0 (Legitimate txn)	3937.6 96.9 96.9	128.2 3.2 **3.2**
1 (Fraud txn)	0.07 0 15.1	0.38 0.01 **84.9**

Logistic Regression

Actual	Predict	
Frequency Percent Row Percent	0	1
0 (Legitimate txn)	3873.4 95.26 95.27	192.3 4.73 **4.73**
1 (Fraud txn)	0.08 0 17.30	0.37 0.01 **82.70**

Tree-Based Model

Actual	Predict	
Frequency Percent Row Percent	0	1
0 (Legitimate txn)	3870.6 94.09 94.10	242.7 5.90 **5.90**
1 (Fraud txn)	0.05 0 11.93	0.39 0.01 **88.07**

CONCLUSION

This paper illustrates a successful example of applying predictive modeling to identify a rare event: a few thousand frauds in millions of transactions. In business and consumer models, neural network model achieved higher hit rate at an acceptable false positive rate as planned. There are several keys to the success: a key strategy of looking for the deviation of fraud transaction behavior from a normal one, generation of transaction-based variables as predictors for the model, and utilization of Enterprise Miner® in data analysis and modeling. The ease-of-use of Enterprise Miner® shortens the development time significantly. In developing a major risk management model, the ease of implementation and integration into the existing infrastructure are also criteria to the selection of an appropriate model.

REFERENCES

SAS® Enterprise Miner™ Software.
Applying Data Mining Techniques. SAS institute Inc.

Author Contact:

Shiao-ping Lu
2669 Greer Road
Palo Alto, CA
650-843-1478
splu@lucidconsult.com