Paper 113-2013

Detecting Cross Channel Fraud using SAS®

Srikar Rayabaram and Krutharth Peravalli, Oklahoma State University, Stillwater, OK

ABSTRACT

In a world where criminals are getting effective in their ability to gain information about a customer of a particular bank, cross channel monitoring and assessment has become very important. As each day passes by, criminals are also getting bolder in terms of engaging beyond a single channel to set in motion the movement of money. In these scenarios, a cross channel review of user activity is very much essential to detect or prevent fraud. In this paper we will analyze data across various channels such as cash, regular wires and FCE wires of a south east bank to detect fraudulent activity along with Money Service Businesses (MSB). Also, we will be creating a predictive model which can be used to predict such activity and discuss how effective the model would have been to detect fraudulent activity in the past.

DATA PREPARATION:

Our analysis window is January 1st 2011 until December 31st 2011. For the analysis, we created two final datasets, one for training the model and the other for testing the model. The training dataset has the information from January 1st 2011 until September 30th 2011 i.e. Q1, Q2, Q3 for 2011. To arrive at the data we need for the analysis and building a model, we had to effectively use a set of files and create the final datasets. In this section, we explain about the list of the files that are used to create the final datasets.

- ***Customer Information***: A flat file was available that has the information of the customers like the customer number, customer name, customer type, w8, country etc. We used the w8 information and tax id information to create a flag called “nra_flag” which says whether a customer is NRA (Non Resident Alien) or not. We created a SAS dataset with the above information. The data needed for training dataset has been created using the customer information until September 30th 2011 and the data needed for test dataset has been created using the information from October 1st 2011 until December 31st 2011.

- ***Deposit Information***: A flat file was available which has the account number, customer number, account ledger balance, type of account etc. We have chosen the above mentioned four columns and created the SAS datasets. We created four SAS datasets, each having customer number, account number and account balance at end of each quarter in 2011. We then merged the information of the first 3 quarters of 2011 to use the same as training data and the Q4 information as Test data.

- ***Cash Transactions***: A set of flat files were available that has the daily cash transaction information for the customers from January 1st 2011 until December 31st 2011. These files have the customer number and transaction date along with the other transaction

information. We created the dataset in such a way that it has total cash amount for a customer in a quarter. We then merged the information of the first 3 quarters of 2011 to arrive at the intermediate cash data set which will be used to create the final training dataset. We used the data of the 4th quarter to create the final test data.

- ***FCE Wire Transactions***: A set of flat files (which are provided by third party to the financial institution) were available that has the daily FCE (Foreign Currency Exchange) wire transaction information for the customers from January 1st 2011 until December 31st 2011. These files have the customer number and transaction date along with the other transaction information. We initially created the SAS dataset that has the account number and total amount of regular wire transactions for an account in a quarter. We then merged the data of the first 3 quarters to arrive at the dataset that will be used to create the final training data set. We used the data of the 4th quarter to create the final test data.

- ***Regular Wire Transactions***: A set of flat files were available that has the daily wire transaction information for the customers from January 1st 2011 until December 31st 2011. These files have the account number and transaction date along with the other transaction information. We created the dataset in such a way that it has total wire amount for a customer in a quarter. We then merged the information of the first 3 quarters of 2011 to arrive at the intermediate wire data set which will be used to create the final training dataset.

- ***Risk Scores***: A set of flat files (which are provided by third party to the financial institution) were available that has the customer number, account number and the risk scores for each customer. These files are monthly batch files and have the risk scores for each customer for a month. We created intermediate SAS data sets that have the latest risk scores for a customer in a quarter. We then merged the data of the first 3 quarters which is used to create the final training dataset and the data of 4th quarter to create the final test dataset.

- ***FINCEN:*** The Fincen data for every quarter in 2011 was downloaded from the "http://fincen.gov/" website. We then merged the FINCEN files with the customer dataset to mark a customer as MSB if that customer has appeared in the FINCEN list. We then merged the data for the first 3 quarters to use the same to create the final training dataset and used the data of 4th quarter to create the final test dataset.

- ***Target Variable:*** We were given a list of High Risk Customers (which was done by a third party for the financial institution based on some business rules) for the company for each quarter. We then merged the data for first 3 quarters to use the same to create the final training dataset. We used the data of 4th quarter to create the final dataset.

Final Training Dataset: The customer, deposit, cash, regular wire, FCE wires, risk scores, MSB datasets which have the data for the first 3 quarters, have been used to create the final training dataset. The process is given below.

Step 1: Since the regular wire transaction dataset has account number, we joined the same with the deposit dataset which has both the account number and customer number. We now have a single dataset that has both the account balance and the total amount of wire transaction for the first 3 quarters of 2011.

Step 2: The dataset created in Step 1 has been joined with the customer dataset on customer number to create a dataset that has the customer, deposit and regular wire data for the first 3 quarters of 2011.

Step 3: The dataset created in Step 2 is joined with the cash dataset on customer number to create a dataset that has the customer, deposit, regular wire and cash data for the first 3 quarters of 2011.

Step 4: The dataset created in Step 3 is joined with the FCE wires dataset on customer number to create a dataset that has the customer, deposit, regular wire, cash and FCE wires data for the first 3 quarters of 2011.

Step 5: The dataset created in Step 4 is joined with the risk scores dataset on customer number to create a dataset that has the customer, deposit, regular wire, cash, FCE wires and risk scores data for the first 3 quarters of 2011.

Step 6: The dataset created in Step 5 is joined with the fincen dataset on customer number to create a dataset that has the customer, deposit, regular wire, cash, FCE wires, risk scores and MSB data for the first 3 quarters of 2011.

Step 7: The dataset created in Step 6 is joined with the high risk dataset on customer number to create a dataset that has the customer, deposit, regular wire, cash, FCE wires, risk scores and high risk customers' data for the first 3 quarters of 2011 to arrive at the final training dataset.

Final Test Dataset: The customer, deposit, cash, regular wire, FCE wires, risk scores, MSB datasets which have the data for the 4th quarter, have been used to create the final test dataset. The process to create the final test dataset is similar to the process that has been used to create the final training dataset.

INITIAL ANALYSIS:

The primary objective is to train the data for the first 3 quarters and predict the high risk customers for the fourth quarter of 2011. The training dataset has 86,159 observations out of which 0.005% observations have target (hr_flag) equal to 1. Hence, we took a stratified sample from the population dataset with sample size as 5000 observations. The final sample dataset has 7078 observations of which 2078 observations have target equal to 1.

Summary statistics on the final training dataset are as follows:

Summary Statistics

Results

The MEANS Procedure

Variable	Mean	Std Dev	Minimum	Maximum	N	Median
nra_flag	0.0499658	0.2178755	0	1.0000000	86159	0
acct_bal	197299.95	3155207.04	0	233197675	86159	5529.31
total_reg_wire	165261.60	4535554.65	0	773719490	86159	0
total_cash	10611.01	549835.42	0	150306493	86159	23.0000000
total_wire	362420.97	6946025.34	0	773719490	86159	0
risk_score	21.3183997	142.1546505	0	6612.00	86159	0
msb	0.0040158	0.0632436	0	1.0000000	86159	0
hr_flag	0.0241182	0.1534170	0	1.0000000	86159	0

Summary statistics of the random sample generated on final dataset are as follows:

Summary Statistics

Results

The MEANS Procedure

Variable	Mean	Std Dev	Minimum	Maximum	N	Median
hr_flag	0.2935858	0.4554366	0	1.0000000	7078	0
nra_flag	0.0536875	0.2254159	0	1.0000000	7078	0
acct_bal	1398091.40	10693970.69	0	233197675	7078	19563.77
total_cash	66765.89	691027.80	0	17811525.59	7078	300.0000000
total_wire	3600240.68	23668325.51	0	773719490	7078	0
risk_score	133.6363379	444.8550075	0	6612.00	7078	0
msb	0.0063577	0.0794871	0	1.0000000	7078	0

Summary statistics of the final testing dataset are as follows:

Summary Statistics

Results

The MEANS Procedure

Variable	Mean	Std Dev	Minimum	Maximum	N	Median
acct_bal	590885.13	22757610.98	0	2477435982	29443	4502.62
total_reg_wire	189065.69	5314452.44	0	715407217	29443	0
total_cash	5963.68	131560.80	0	13933849.86	29443	0
total_wire	231715.76	8014449.48	0	1210198849	29443	0
risk_score	20.6064599	114.7409594	0	7438.00	29443	0
msb	0.0038719	0.0621050	0	1.0000000	29443	0
hr_flag	0.0016303	0.0403443	0	1.0000000	29443	0
nra_flag	0.0516252	0.2212729	0	1.0000000	29443	0

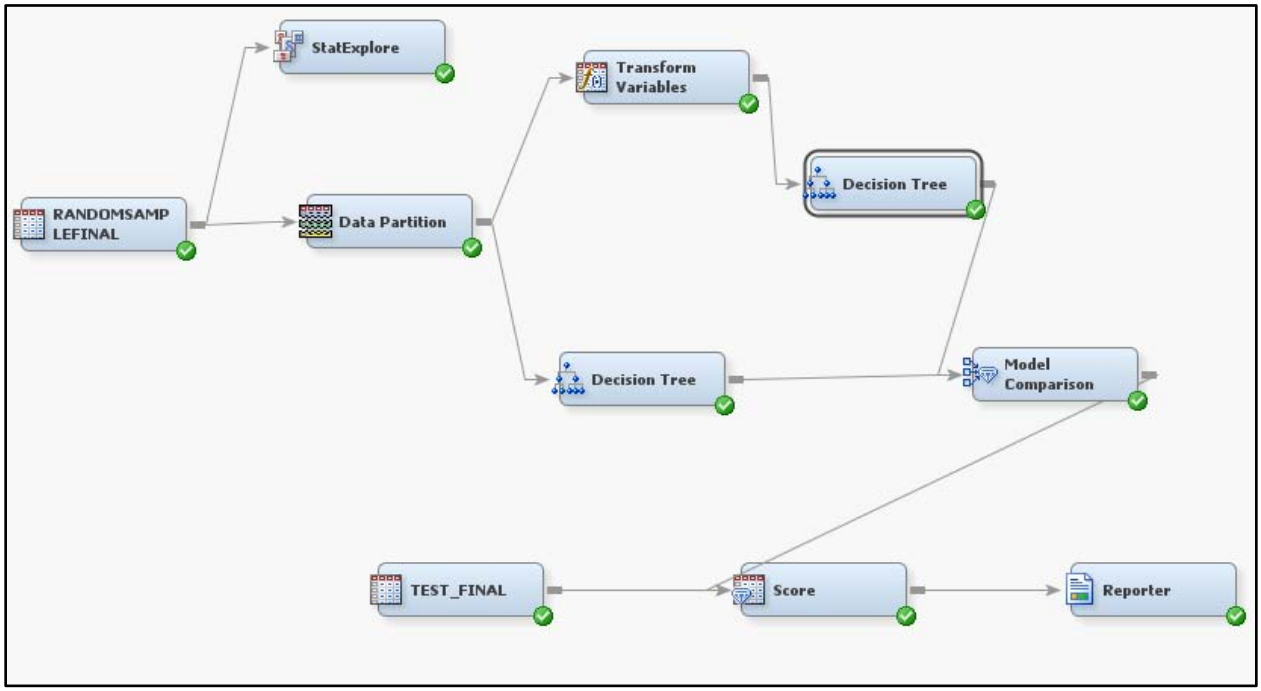
Now we try to see whether there is any variable which is strongly correlated with the target variable (hr_flag). The correlations of the input variables with the target variables are as follows.

Pearson Correlation Coefficients, N = 86159 Prob > |r| under H0: Rho=0	hr_flag
nra_flag	-0.00168 0.6226
acct_bal	0.21436 <.0001
total_cash	0.05946 <.0001
total_wire	0.26435 <.0001
risk_score	0.44942 <.0001
msb	0.02232 <.0001

From the Pearson correlations, we can see that the variable risk_score has a strong positive relationship with the target variable. The variables acct_bal, total_wire are weakly (but positively) related with target variable.

Model:

The best model for predicting the possibility of a customer being a high risk customer was the decision tree model. Even more surprisingly, the entropy decision tree model with the non-transformed inputs had better misclassification rates.



Interpretation of Results:

As a decision tree model without the transformed inputs did a better job in predicting the high risk customers, the classification of a customer as a high risk customer or not is based on the different types of transactions happening across various channels like wire transfers, cash transfers etc., Further looking into the results, we can infer from the decision tree output that our assumption on how the classification of a customer as a high risk customer is done is accurate to a certain extent.

Variable Importance

Obs	NAME	LABEL	NRULES	IMPORTANCE
1	total_wire		2	1.00000
2	total_cash		3	0.23706
3	acct_bal		3	0.18000
4	type		2	0.13382
5	risk_score		2	0.09703
6	nra_flag		1	0.07438

The top 3 important variables in predicting the target variable are total_wire, total_cash and acct_bal. Based on the event classification table, one can easily infer that the misclassification rate for this model is nearly just above 4.4%. The entropy tree with transformed inputs had a misclassification rate of nearly 6%.

Classification Table

Data Role=TRAIN Target Variable=hr_flag

Target	Outcome	Target Percentage	Outcome Percentage	Frequency Count	Total Percentage
0	0	96.1933	97.5400	4877	68.9036
1	0	3.8067	9.2878	193	2.7268
0	1	6.1255	2.4600	123	1.7378
1	1	93.8745	90.7122	1885	26.6318

When the model was deployed on the final test dataset (quarter 4 data), the misclassification rate was 3.1% when compared to the actual quarter 4 results. The classification table for the final testing data is as follows.

Target	Outcome	Count	%age
0	0	28479	96.72588
0	1	916	3.111096
1	0	11	0.03736
1	1	37	0.125667

Also, from the English rules developed by the decision tree, the following can be inferred.

1) If the amount exchanged through wire transfer is more than 454,741$ then the probability of that customer being a high risk customer is very high.

```
*-------------------------------------------------------------*
 Node = 7
*-------------------------------------------------------------*
if total_wire >= 454741 or MISSING
then
 Tree Node Identifier   = 7
 Number of Observations = 1763
 Predicted: hr_flag=0   = 0.06
 Predicted: hr_flag=1   = 0.94
```

2) If the amount exchanged through wire transactions is less than 144,682$ and the amount exchanged through cash transactions is greater than 413,604$, then the probability of the customer being a high risk customer is very high.

```
*-------------------------------------------------------------*
 Node = 11
*-------------------------------------------------------------*
if total_wire < 144682 or MISSING
AND total_cash >= 413604
then
 Tree Node Identifier   = 11
 Number of Observations = 46
 Predicted: hr_flag=0   = 0.02
 Predicted: hr_flag=1   = 0.98
```

Conclusion:

The datasets provided to us are aggregated to a quarterly level. Based on the data which was spread across three quarters we were able to identify or detect high risk customers for the fourth quarter. Also, based on the analysis and the predictive model results it is safe to conclude that the third party organization which identifies the high risk customer for this bank uses the transactions happening across various channels like wire transfer, cash transfer etc., to identify the same. The model which has been developed will help this bank in identifying fraudulent transactions only on a quarterly basis. This is just the first half of the problem. It would be really beneficial for the organization if the transactions are monitored on a daily basis. Finally, to fight against these new age outlaws, organizations need to invest more on analytics. After all, as mentioned by Arthur C. Nielsen, the price of light is less than the cost of darkness.

APPENDIX:

The list of variables used in the final datasets and their descriptions is provided in the below table.

Variable	Usage	Data Type	Description
Cust_no	Input	Numeric	Customer Number
Cust_name	Input	Character	Customer Name
Type			Customer Type, if type='I' then Individual Customer else if type='O' then Business/Organization customer.
Country_code	Input	Character	Country code of the customer.
Country	Input	Character	Country Description of the customer.
Nra_flag	Input	Numeric	If nra_flag=1, then the customer is nra (non resident alien) else if nra_flag=0 then the customer is not nra.
Acct_bal	Input	Numeric	Total account balance of a customer in a quarter.
Total_reg_wire	Input	Numeric	Total amount of regular wire transactions by a customer in a quarter.
Total_cash	Input	Numeric	Total amount of cash transactions by a customer in a quarter.
Total_wire	Input	Numeric	Total amount of FCE (Foreign Currency Exchange) wire transactions by a customer in a quarter.
Risk_score	Input	Numeric	Risk Score for a customer in a quarter.
Msb	Input	Numeric	If msb=1 then the business/Organization has appeared in the FINCEN MSB list else if msb=0 then the business/Organization has not appeared in the FINCEN MSB list.
Hr_flag	Target	Numeric	If hr_flag=1 then the customer is high risk customer else if hr_flag=0 then the customer is not high risk customer.

References:

Enterprise wide Fraud Management - Ellen Joyner, SAS Institute Inc. Cary, NC, USA. Paper 029-2011, SAS Global forum 2011.

Contact Information:

Your comments and questions are valued and encouraged. Contact the authors at:

Srikar Rayabaram, Email: rayabar@okstate.edu

Srikar Rayabaram is Master's student in Management Information Systems at Oklahoma State University. He is a BASE SAS® 9 and a certified SAS® predictive modeler using Enterprise Miner 6. He has also received his SAS® and OSU Data Mining Certificate in December 2012.

Krutharth Peravalli, Email: kruthar@okstate.edu

Krutharth Peravalli is Master's student in Management Information Systems at Oklahoma State University. He is a BASE SAS® 9 and a certified SAS® predictive modeler using Enterprise Miner 6. He has also received his SAS® and OSU Data Mining Certificate in December 2012.