Expert Systems with Applications 39 (2012) 12650–12657

Contents lists available at SciVerse ScienceDirect

Expert Systems with Applications

journal homepage: www.elsevier.com/locate/eswa

Employing transaction aggregation strategy to detect credit card fraud

Sanjeev Jha [a,*], Montserrat Guillen [b,1], J. Christopher Westland [c,2]

[a] Department of Decision Sciences, Whittemore School of Business and Economics, University of New Hampshire, McConnell Hall, Durham, New Hampshire 03824-3593, USA
[b] Department of Econometrics, Riskcenter-IREA, University of Barcelona, Diagonal, 690, 08034 Barcelona, Spain
[c] Department of Information & Decision Sciences (MC 294), Room 2400, University Hall, University of Illinois, Chicago, 601 S. Morgan Street, Chicago, IL 60607-7124, USA

ARTICLE INFO

Keywords:
Fraud detection
Predictive modeling
Logistic regression

ABSTRACT

Credit card fraud costs consumers and the financial industry billions of dollars annually. However, there is a dearth of published literature on credit card fraud detection. In this study we employed transaction aggregation strategy to detect credit card fraud. We aggregated transactions to capture consumer buying behavior prior to each transaction and used these aggregations for model estimation to identify fraudulent transactions. We use real-life data of credit card transactions from an international credit card operation for transaction aggregation and model estimation.

© 2012 Elsevier Ltd. All rights reserved.

1. Introduction

Credit card fraud costs consumers and the financial industry billions of dollars annually (Chan, Fan, Prodromidis, & Stolfo, 1999; Chen, Chen, & Lin, 2006). The reported loss due to online fraud for the year 2008 was $4 billion, an increase of 11% on year 2007 loss of $3.6 billion (Leggatt, 2008). Credit card transactions, as a share of payment system, have been growing worldwide along with credit card fraud. Moreover, credit card fraud funds other criminal activities, including terrorism in ways that may be difficult to track and prevent (Everett, 2009). As fraud detection has steadily evolved, perpetrators have become more sophisticated in tandem with these improvements (Bolton & Hand, 2002). The audit of credit card fraud is an ongoing 'arms-race' that requires constant innovation on the part of card issuers.

However, there are various obstacles to this innovation. For example, academicians have difficulty in getting credit card transactions datasets leading to less academic research and also not much of proprietary detection techniques get discussed in public lest fraudsters should gain knowledge and evade detection (Leonard, 1993). There is a dearth of published literature on credit card fraud detection, which makes exchange of ideas and possible innovation in fraud detection difficult (Bolton & Hand, 2002). One difficulty with analysis of credit card fraud is that perpetrators do not usually carry on a single fraudulent transaction. Analyzing fraud from the perspective of a "one by one" transaction omits the idea of clustering that is inherent of credit card fraud actions. Perpetrators usually produce a group of fraudulent transactions. We argue that analyzing the aggregated behavior is essential to improve credit card fraud detection rates.

In this study we employ transaction aggregation strategy (Krivko, 2010; Whitrow, Hand, Juszczak, Weston, & Adams, 2009) to create variables for the estimation of a logistic regression model to attempt to detect (and thus control and prosecute) credit card fraud. We demonstrate the efficacy of aggregating transactions to capture consumer buying behavior prior to each transaction. The underlying rationale is that the buying behavior of fraudulent and legitimate transactions is different. This difference gets captured in aggregated transactions and can be used for identification of fraudulent transactions. We use real-life data of transactions from an international credit card operation for aggregating transactions and then use them for model estimation.

A general definition of 'fraud' may be somewhat elusive, as new methods of fraud appear with regularity. For the purpose of this study, fraudulent transactions are specifically defined by the institutional auditors as those that caused an unlawful transfer of funds from the bank sponsoring the credit cards. These transactions were observed to be fraudulent ex post.

The remainder of this paper is organized as follows. In the next section we discuss credit card fraud and detection methods. In section 3, we discuss the dataset source, primary attributes, and creation of derived attributes using these primary attributes. In section 4, we discuss the estimation method and present a standard logit model. Thereafter, we present our results, discussion, and conclusions of our study.

* Corresponding author. Tel.: +1 603 862 0314; fax: +1 603 862 3383.
E-mail addresses: sanjeev.jha@unh.edu (S. Jha), mguillen@ub.edu (M. Guillen), westland@uic.edu (J. Christopher Westland).
[1] Tel.: +34 93 4037039; fax: +34 93 4021821.
[2] Tel.: +1 312 996 2323; fax: +1 312 413 0385.

0957-4174/$ - see front matter © 2012 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.eswa.2012.05.018

2. Background

2.1. Credit card fraud

Credit card frauds can be committed in a number of ways (Blunt & Hand, 2000). However, credit card frauds have been classified into two broad categories: application and behavioral frauds (Bolton & Hand, 2001). Application fraud occurs when fraudsters obtain new cards from issuing companies and are of two types. In the first type, perpetrators obtain new cards from issuing companies using other people's information and keep using the cards with the stolen identity until fraud is detected. In the second type, perpetrators seek new credit cards using false personal information with the intention of never repaying their purchases (Bolton & Hand, 2002). Financial institutions have traditionally used credit scores to deny issue of credit cards to individuals likely to default payments either because they do not have sufficient income or because they fit the profile of those likely to commit fraud (Hand & Henley, 1997). Also, financial institutions use various models to monitor purchase behavior over time to detect cards obtained using false information. For example, a first time card holder who reaches his or her credit limit within a few days of issuance of a card or exhibits similar unusual purchase behavior may raise alarm. Researchers have employed case-based reasoning approaches to detect credit card application fraud (Wheeler & Aitken, 2000).

Behavioral frauds are of four types: mail theft, stolen/lost card, counterfeit card, and 'card holder not present' fraud. Mail theft fraud occurs when fraudsters intercept credit cards in mails before they reach cardholders. Stolen/lost card fraud happens when fraudsters get hold of credit cards through theft (for example of a purse or wallet). In case of counterfeit card fraud, like the previous two kinds of fraud, a physical card is used to commit fraud. However, the difference is that perpetrators pilfer card information in order to create a physical counterfeit card. For 'card holder not present' fraud, unlike the other three types of fraud, a physical card is not necessary. 'Card holder not present' fraud is done remotely and perpetrators are not present physically at a merchant's premises. Card details are enough to carry out a transaction (Bolton & Hand, 2002). Since transactions are carried out remotely, perpetrators do not have to sign for purchases or to physically swipe credit cards or even show proof of identification. Thus perpetrators carry out fraudulent transactions in complete anonymity. These four types of behavioral fraud represent a very high proportion of losses (Bolton & Hand, 2002). In this study, we investigate behavioral credit card frauds.

Williams (2007) chronicled the manner in which credit card fraud has evolved over the years. In the 1970s, stolen cards and forgery were the most prevalent type of credit card fraud, where physical cards were stolen and used. Later, mail-order-telephone-order fraud became common in the 1980s and 1990s. Now, credit card fraud has moved to the Internet, which provides the anonymity, reach, and speed to commit fraud across the world.

2.2. Fraud detection methods

Financial institutions and merchants fight against fraud at two levels: fraud prevention and fraud detection (Bolton & Hand, 2001). Fraud prevention pertains to all activities and practices engaged in stopping fraud from happening in the first place. An example of fraud prevention is the practice of credit card activation before its first use to prevent theft of credit cards from surface mails. Internet security systems for credit card transactions are another example of fraud prevention. Pin numbers for debit cards are another example of fraud prevention, as perpetrators need to know the pin number and have physical possession of the debit cards in order to withdraw money from ATMs.

Fraud detection, on the other hand, pertains to practices and systems to quickly detect fraudulent transactions as soon as these transactions take place (Bolton & Hand, 2001). The sooner fraudulent transactions are detected the more losses can be avoided by stopping transactions made with fraudulent credit cards. Fraud detection is a continuous activity as there is no way to know if fraud prevention has failed and which transactions are fraudulent. Statistical fraud detection methods have been classified into two broad categories: 'supervised' and 'unsupervised' (Bolton & Hand, 2001).

In supervised statistical methods, estimated statistical models are used to discriminate between fraudulent and non-fraudulent purchase behavior to classify new observations into an appropriate class: fraudulent or non-fraudulent transaction (Bolton & Hand, 2001). The performance of models is assessed by measuring their accuracy in correctly classifying new observations as fraudulent or non-fraudulent. Supervised statistical methods have three important characteristics. First, supervised statistical methods require samples of both classes, fraudulent and legitimate observations, as models are trained based on examples of observations in both classes. Second, supervised statistical models can only detect frauds that have occurred and have been detected previously (Bolton & Hand, 2001). This method cannot detect new kinds of fraud. Hence, the training datasets should have all kinds of fraudulent transactions for appropriate performance of statistical models, and the sample classes have to be periodically updated to include newer kinds of fraud.

Previous research on credit card frauds employing supervised methods can be divided into three categories (Bolton & Hand, 2002): traditional statistical classification methods, rule-based methods, and recent development of power tools. Examples of traditional statistical classification methods are linear discriminant analysis and logistic regression (Hand, 1981; McLachlan, 1992). Rule-based methods, such as tree-based algorithms (Breiman, Friedman, Olshen, & Stone, 1984; Quinlan, 1993) are supervised learning algorithms that use rules of *If* (fulfills certain conditions) ... *Then* (appropriate category). Examples of recent sophisticated "power tool" methods of classification are neural networks (Hand, 1997; Quah & Sriganesh, 2008; Ripley, 1996; Webb, 1999), SVMs (Whitrow et al., 2009), and Random Forest (Whitrow et al., 2009). Kou, Chang-Tien, Sirwongwattana, and Huang (2004) provide a summary of fraud detection techniques used in past research. Phua, Lee, Smith, and Gayler (2005) have done comprehensive survey of data mining based fraud detection research. Bose (2006) has reviewed existing intelligent technologies for managing fraud and identity theft.

Although prior studies have proposed different techniques and algorithms for credit card fraud detection, a number of studies were done in an experimental setup and very few studies used real credit card data. In this research, we use a dataset of real-life credit card transactions to estimate a supervised statistical model.

Unsupervised methods attempt to detect unusual observations, such as customers, transactions, or accounts whose behavior may be different from the norm. These unusual observations, different from the baseline normal behavior, are identified for closer examination and subsequent classification. Unlike supervised methods, unsupervised methods do not require samples of fraudulent and legitimate transactions. Hence, unsupervised methods may find use in situations where there is no prior knowledge of classes of observations. The other advantage of unsupervised methods over supervised method is that previously undiscovered frauds can be detected, while supervised methods can only be trained to detect the kinds of frauds in historical databases. However, unsupervised methods, compared to supervised methods, have been less popular in fraud detection and have not received much attention in fraud detection literature (Bolton & Hand, 2001).

3. Data

Our study uses the dataset of Paasch (2007), which was obtained from an international credit card operation. Paasch (2007) used Artificial Neural Networks (ANN) tuned by Genetic Algorithms (GAs) to detect fraud. This dataset has 13 months, from January 2006 through January 2007, of about 50 million (49,858,600 transactions) credit card transactions on about one million (1,167,757 credit cards) credit cards. We have named this panel dataset of all transactions in this period as dataset A (Fig. 1). A much smaller subset of this large dataset is dataset B, which has 2,420 known fraudulent transactions made with 506 credit cards.

We expected fraudulent transactions to be different from legitimate transactions. Indeed, our descriptive statistics of dataset B showed prevalence of frauds in only a few transaction categories: retail purchases (94.67%), check-item (4.54%) and non-directed payments (0.50%) whereas the full dataset A had more heterogeneity of transaction categories, with retail purchases accounting for only 45% of total transactions (Table 1). Based on these observations, we partitioned the dataset A to include only the three transaction categories found in dataset B. The reduced dataset A had 31,671,185 transactions.

To estimate our model we needed fraudulent and legitimate transactions. Since the number of cases of observed fraudulent transactions was much less than legitimate transactions, we oversampled fraud cases for this study (Artis, Ayuso, & Guillen, 2002; Duman & Ozcelik, 2011). First, from the reduced dataset A, we created dataset C of all transactions with 506 fraudulent credit cards. Dataset C had 37,280 transactions, out of which 2420 were fraudulent transactions and 34,860 were legitimate transactions (Fig. 1). Then, excluding the transactions in dataset C (which included dataset B transactions), from the remaining transactions in reduced dataset A we pulled up transactions on 506 randomly selected credit cards. This comparable dataset had 12,679 legitimate transactions and we named this dataset as D. In sum, our test dataset had 2420 fraudulent transactions (dataset B) and 12,679 legitimate transactions (dataset D).

Table 1
Percentage of credit card transactions by transaction types.

Transaction types	Dataset A	Dataset B
Retail purchase	48.65	94.67
Disputed transaction	15.58	0.00
Non directed payment	14.15	0.50
Retail payment	8.85	0.00
Miscellaneous fees	4.11	0.00
Transaction code	3.91	0.00
Cash-write-off-debt	1.30	0.00
Cash-adv-per-fee	0.62	0.00
Check-item	0.63	4.54
Retail-adjust	0.01	0.00
Others	2.19	0.29
Total	100.00	100.00

3.1. Primary attributes in datasets

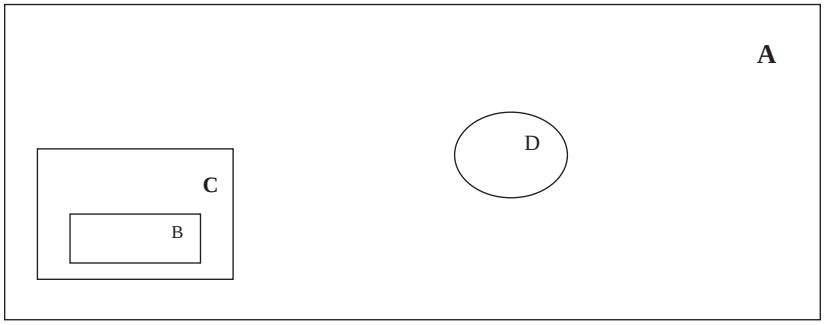
For the purpose of this study, we name the attributes of credit card transactions available in the above datasets as *primary attributes*. We present these attributes in Table 2. Posting date attribute is the date of posting of transactions to the accounts. Account number attribute is the 16 digit credit card number of each transaction. Transaction type attribute categorizes transactions into types of transactions like retail purchase, non-directed payments etc. Currency attribute provides the short code for the currency used to perform a transaction. Merchant code is the category code for the merchant for a given transaction. Foreign currency transaction, a binary attribute, flags a given transaction into whether a transaction is in foreign currency or not. Transaction date attribute is the date of a transaction. Merchant name, merchant city, and merchant country attributes describe the merchants of respective transactions. The acquirer reference code is a unique identifier for each transaction. E-commerce flag is a binary variable indicating if a transaction was an e-commerce transaction.

There are two quantitative attributes of transactions in the dataset: foreign and local currency amount. Foreign currency amount is the amount of transaction made with a foreign merchant. Local currency amount is the amount of transaction in the local currency of the country where the card was issued.

We make a note here that although the dataset had transaction and posting dates of each transaction, there is no time stamp attribute available. In other words, in a given business day, we have no way to know the sequence of each transaction on a credit card. This is one of the limitations of this study and we discuss this again in sections below. Also, the other limitation of the dataset is the non-availability of demographic details of the credit card holders. Availability of demographic data would have enriched our analyses and findings. However, we appreciate the reason for reluctance in providing the demographic details of credit card holders by the financial institution.

3.2. Derived attributes used in testing

Whitrow et al. (2009) in their recent article advocate transaction aggregation as a strategy for credit card fraud detection. The authors argue that it is not very practical to employ all credit card



Dataset A: All transactions	49,858,600 transactions
Dataset B: Fraud Dataset	2,420 transactions
Dataset C: All transactions with Fraudulent Credit Cards	37,280 transactions
Dataset D: Transactions with random sample of 506 credit cards	12,679 transactions

Fig. 1. Dataset description.

Table 2
Primary attributes in datasets.

Attribute name	Description
Posting date	Date when transaction was posted to the accounts
Account number	Credit card number
Transaction type	Transaction types, such as cash advance and retail purchase
Currency	Short code of the currency the transaction was originally performed in
Merchant code	Merchant category code
Foreign Txn	Flagging whether this Is a foreign currency transaction
Transaction date	Date when the transaction was actually performed
Merchant name	Name of merchant
Merchant city	City of merchant
Merchant country	Country of merchant
Acquirer reference	The reference code for the transactions
Ecommerce	Flag if this was an Internet transaction
Foreign Txn amt	Foreign currency amount, the amount when transaction was made with a foreign merchant
Local Txn amt	The amount in the local currency of the country where the card of issued

transactions to the fraud detection system because of high dimensionality of credit card transactions data and also because of the heterogeneity in credit card transactions. The authors suggested transaction aggregation strategy as a way of aggregating information over a succession of transactions. In this study we have followed this strategy and have aggregated consumer buying behavior prior to each transaction. We named these aggregations as derived attributes and compute them from primary attributes available in the credit card transaction datasets, as discussed in the previous section.

As mentioned in the above section, there were only two numerical attributes in our dataset: foreign currency amount and local currency amount. The other attributes were categorical. Hence, similar to Paasch (2007), we created derived numerical attributes, aggregating past transactions for each transaction in the dataset to put each credit card transaction into the historical context of past shopping behavior.

We would like to mention here that there are two important issues to keep in mind while creating derived attributes: choice of primary attributes and the length of aggregation. In this research, although, we have been guided by past research, the possible combinations of primary attributes and time periods are infinite. For example, a possible derived attribute can be the number of transactions with a specific merchant. The aggregation of the number of transactions can be over a day, a week, a month, three months, or any other time period. Researchers or financial institutions have to be judicious about the periods of aggregation, because length of aggregation period can have an impact on the performance of statistical models (Whitrow et al., 2009). While aggregation over a very small period may not capture sufficient history of consumer spending pattern, aggregation over a longer duration may capture too much of noise and hide the relevant pattern distinguishable over shorter period. There is no limit to experimentation with the choice of time periods. Also, the choice of time periods could vary from one derived attribute to the other.

In this study, we have created derived attributes by aggregating over a day, a month, and three months. We fully understand that the aggregation period could have been different from what we have done, but as we mentioned above, there are infinite ways of aggregation. However, for this study we have created 14 derived attributes using different primary attributes and time periods (Table 3). As we noted in the previous section, since we do not have time stamping data, all transactions with a credit card on any day had the same value for all derived attributes.

Txn amount over month: Average spending per transaction over a 30-day period on all transactions till this transaction. We computed the total amount spent with a credit card during past 30 days prior to a transaction and divided it by the number of transactions to get the average amount spent.

Average over 3 months: Average amount spent over the course of 1 week during past 3 months. For this attribute, we computed the total amount spent with a credit card during past 90 days prior to a transaction, and then divided it by 12 to get the average weekly spent over three months.

Average daily over month: Average spending per day over past 30 days before this transaction. We calculated the total amount spent with a credit card during past 30 days prior to a transaction and divided it by 30 to compute the average daily spent over a month prior to a transaction.

Amount merchant type over month: Average spending per day on a merchant type (based on merchant code) over a 30-day period for each transaction. In this case, we first computed the total amount spent with a credit card on a specific merchant type during past 30 days prior to a transaction and then divided this sum by 30 to get the average money spent with a specific merchant type over a month prior to a transaction.

Number merchant type over month: Total number of transactions with the same merchant over a period of 30 days before a given transaction. For this attribute, we computed the total number of transactions with a credit card with a specific merchant type during past 30 days prior to a transaction.

Amount merchant type over 3 months: Average weekly spending on a merchant type during past 3 months before a given transaction. For this attribute, we computed the total amount spent with a credit card on a specific merchant type during past 90 days prior to a transaction, and then divided it by 12 to get the average weekly amount spent over three months on that merchant type.

Amount same day: Total amount spent with a credit card on the day of a given transaction. Here, for each transaction we computed the total amount spent in a day with that credit card.

Number same day: Total number of transactions on the day of a given transaction. For this attribute, we computed the total number of transactions in a day with that credit card.

Amount same merchant: Average amount per day spent over a 30 day period on all transactions up to this transaction on the same merchant as this transaction. In this case, we computed the average amount spent on the same merchant during last month.

Number same merchant: Total number of transactions with the same merchant during last month. For this attribute, we computed the total number of transaction with the same merchant during last 30 days.

Amount currency type over month: Average amount spent over a 30-day period on all transactions up to this transaction with the same currency. For this attribute, we first computed the total amount spent with a credit card on a specific currency type during past 30 days prior to a transaction and then divided this sum by 30 to get the average money spent with a specific currency type over a month prior to a transaction.

Number currency type over month: Total number of transactions with the same currency type during past 30 days. For this attribute, we computed the total number of transactions with a credit card with a specific currency type during past 30 days prior to a transaction.

Amount same country over month: Average amount spent over a 30-day period on all transactions up to this transaction in the same country. For this attribute, we first computed the total amount

Table 3
Derived attributes used in models.

Short name	Description
Txn amount over month	Average amount spent per transaction over a month on all transactions up to this transaction
Average over 3 months	Average amount spent over the course of 1 week during past 3 months
Average daily over month	Average amount spent per day over the past 30 days
Amount merchant type over month	Average amount per day spent over a 30 day period on all transactions up to this one on the same merchant type as this transaction
Number merchant type over month	Total number of transactions with same merchant type during past 30 days
Amount merchant type over 3 months	Average amount spent over the course of 1 week during past 3 months on same merchant type as this transaction
Amount same day	Total amount spent on the same day up to this transaction
Number same day	Total number of transactions on the same day up to this transaction
Amount same merchant	Average amount per day spent over a 30 day period on all transactions up to this transaction on the same merchant as this transaction.
Number same merchant	Total number of transactions with the same merchant during last month
Amount currency type over month	Average amount per day spent over a 30 day period on all transactions up to this one on the same currency type as this transaction
Number currency type over month	Total number of transactions in the same currency during past 30 days
Amount same country over month	Average amount spent over a 30 day period on all transactions up to this one on the same country as this transaction
Number same country over month	Total number of transactions in the same country during past 30 days before this transaction

spent with a credit card in a specific country during past 30 days prior to a transaction and then divided this sum by 30 to get the average money spent in a specific country over a month prior to a transaction.

Number same country over month: Total number of transactions in the same country during past 30 days before this transaction. In this case, we computed the total number of transactions with a credit card in a specific country during past 30 days prior to a transaction.

4. Model

Binary choice models are appropriate when dependent variable is categorical. Binary choice models have been used in studying frauds in insurance, food stamp programs, and so forth (Artis et al., 2002; Bollinger & David, 1997; Hausman, Abrevaya, & Scott-Morton, 1998; Poterba & Summers, 1995). Caudill, Ayuso, and Guillen (2005) argue that identifying fraudulent claims is similar in nature to several other problems in real life including medical and epidemiological problems. Magder and Hughes (1997) estimated a logistic regression model to predict cessation of smoking. Jin, Rejesus, and Little (2005) used binary choice models in the case of insurance frauds to predict the likelihood of a claim being fraudulent. Investigators use the estimated probabilities to flag individuals that are more likely to submit a fraudulent claim. In this study, our dependent variable *fraud* is binary and we estimate a logistic regression model to predict fraud using primary and derived attributes as independent variables.

5. Results

We first carried out the descriptive statistics of datasets B and D (Table 4). As mentioned in Section 3, Dataset B has 2420 fraudulent transactions, while Dataset D has 12,679 legitimate transactions from 506 randomly chosen cards. A comparison of the average values of derived and primary attributes for the two datasets reveal differences in purchase behavior between fraudulent and legitimate transactions. For example, for the derived attribute *Txn amount over month*, dataset B with just fraudulent transactions has an average of 3590, much higher than the average of 1292 for dataset D with observed legitimate transactions. Similarly, average amount aggregated over a day, month, same merchant type, same merchant, currency type, and same country were higher for fraudulent transactions compared to legitimate transactions.

Table 4
Summary statistics.

Variable	Dataset B (2420 Fraudulent Txn)		Dataset D (12,679 Legitimate Txn)	
	Mean	Std. Dev	Mean	Std. Dev
Txn amount over month	3590	7547	1292	2299
Average over 3 months	5316	13516	3353	6618
Average daily over month	1301	3103	541	1039
Amount merchant type over month	298	521	103	289
Number merchant type over month	3.68	4.58	2.79	2.9
Amount merchant type over 3 months	785	1352	504	1411
Amount same day	7817	13558	1686	5283
Number same day	2.81	2.44	1.26	0.67
Amount same merchant	5019	11249	1385	4553
Number same merchant	1.3	0.87	1.05	0.26
Amount currency type over month	747	1308	323	597
Number currency type over month	8.3	7.9	7.98	7.76
Amount same country over month	655	847	258	534
Number same country over month	8.1	7.81	8.12	10.56
Fraud	1	0	0	0
Foreign Txn Amt	137421	1242628	3474	17588
Local Txn Amt	3857	7798	1302	4363
Foreign Txn	0.68	0.46	0.03	0.18
Ecommerce	0.1	0.3	0	0.05

Furthermore, average number of transactions same day, with same merchant type, same merchant, and same currency were higher for fraudulent transactions compared to legitimate transactions. However, we did not find any difference in the average number of transactions in the same country during past 30 days between fraudulent and legitimate transactions.

Additionally, the average amounts for foreign and local transactions were higher for fraudulent transactions compared to legitimate transactions. Also, it was interesting to find that 68% of fraudulent transactions were foreign transactions compared to only 3% of legitimate transactions (Table 4). Internet transactions were also higher for fraudulent transactions (10%) compared with legitimate transactions. In summary, descriptive statistics of aggregated purchase behavior reveal that fraudulent transactions have higher transaction amount, higher number of transactions, higher amount of local and foreign transaction, and higher number of foreign transactions compared to legitimate transactions.

Subsequently, we ran stepwise regression to select the logistic regressors. The dependent variable, type of credit card transaction, was binary, the type of credit card transaction, legitimate or fraudulent. Legitimate transactions were labeled as 0 (zero) and

Table 5
Estimation results for the logit model.

	Logit model	
	Coefficients	*P* value
Constant	−4.5539	<0.000
Txn amount over month	0.000147	<0.000
Average over 3 months	−0.00006	<0.000
Amount merchant type over month	0.00583	<0.000
Number merchant type over month	0.0378	0.0093
Amount merchant type over 3 months	−0.00234	<0.000
Number same day	0.9507	<0.000
Number same merchant	−0.2048	0.0279
Number currency type over month	0.0858	<0.000
Amount same country over month	0.000722	<0.000
Number same country over month	−0.073	<0.000
Foreign Txn	4.0561	<0.000
Ecommerce	2.2738	<0.000
Log-likelihood	−2782.788	

Restricted log-likelihood = −6645.469.
*Likelihood ratio test = 7725.36; pseudo-R^2 = 58.13%.

fraudulent transactions being the outcome of interest, were labeled as 1 (one). Independent variables were all of the derived attributes in Table 3 and the primary attributes *Foreign Txn Amt, Local Txn Amt, Foreign Txn, and Ecommerce* from Table 2.

The estimation result for logit model is presented in Table 5. The likelihood ratio test for the logit model is 7725.36 with 12 degrees of freedom. This indicates a significant improvement in the model with explanatory variables compared to restricted model with just the constant term. The pseudo-R^2 value of 58.13% indicates improvement in the log likelihood of logit model relative to the baseline model with just the constant term.

As per our expectation, the coefficient for the variable *Txn amount over month* was positive and significant. This indicates that the higher the average spending per transaction over a month before a transaction, the more likely it is that this transaction is fraudulent. The exponentiated coefficient of over 1.0 shows that the odds of the transaction being fraudulent increases with a unit increase in the variable (see Table 5).

However, we found that the estimated coefficient of the derived attribute *Average over 3 months* was negative and significant. This indicates that the increase in average spending over a period of 3 months before a given transaction lowers the probability of the transaction to be fraudulent. The exponentiated coefficient is lower than 1, indicating that the odds of the transaction being fraudulent decreases with a unit increase in the variable. This result is in contrast to the previous variable (*Txn amount over month*), which lends credence to the view that the perpetrators try to maximize frauds by using credit cards as much as possible within a short period of time before the frauds are found and accounts get closed. Hence, a unit increase in the average transaction amount over a month before transaction seem to have higher probability of fraud compared to an increase over 3 months, the rationale being that the perpetrators exhaust all possibilities of fraud within short period of time and therefore probability of fraud is likely to be higher for a unit increase in average spending over a period of a month compared to 3 months before a transaction.

The coefficient corresponding to the derived variable *Amount merchant type over month*, average monthly spending on all transactions up to a given transaction with the same merchant type, was positive and significant. This shows that an increase in the average spending on a merchant type over a month prior to the transaction increases the probability of the transaction being fraudulent. Similarly, we also found that the influence of the variable *Number merchant type over month*, total number of transactions with the same merchant type with a credit card over prior month, was positive and significant. This shows the preference of perpetrators for specific merchant types with whom they prefer to commit frauds.

However, in line with the findings above, average amount spent per week over past three months on transactions with the same merchant type, derived variable *Amount merchant type over 3 months*, was negative and significant. This again indicates that perpetrators commit frauds with a given merchant type over a short period of time before suspicion of frauds are raised and card gets withdrawn. The exponentiated coefficient below 1 indicates that the odds of probability of a transaction being fraudulent get reduced with increase in the value of this derived variable.

The estimate of the parameter for the total number of transactions same day, *Number same day*, was positive and significant. The value of the exponentiated coefficient was 2.87. This shows that the odds of a transaction being fraudulent increase substantially with an increase in number of transactions same day. In other words, the higher the number of transactions with a credit card same day, the higher the probability of the next transaction being fraudulent. This seems probable and agrees with our discussion above that perpetrators attempt to use stolen credit cards quickly to maximize fraud amounts.

However, the coefficient of the derived variable *Number same merchant*, the number of transactions same day with the same merchant with a credit card, was negative and significant. The exponentiated coefficient for this derived variable was 0.775, which shows that the odds of a transaction being fraudulent decreases with increase in the value of this variable. This is in contrast to the previous result, but is plausible because the perpetrators may not like to risk doing a number of transactions with a credit card with the same merchant on a given day.

The coefficient for the derived variable *Number currency type over month* was positive and significant. The exponentiated coefficient was greater than 1. This indicates that the higher is the number of transactions in the same currency over a month, the higher the chances of a transaction being fraudulent.

The other expected result is the positive and significant coefficient of derived variable *Amount same country over month*, average monthly amount spent on a credit card in the same country. This indicates that the higher the average monthly amount spent in the same country, the more likely the next transaction in the same country will be fraudulent.

Interestingly, contrary to our expectations, we found the coefficient of the derived variable *Number same country over month*, the number of transactions in the same country during past month, to be negative. This shows that perpetrators may not prefer to do many transactions in the same country. However, as discussed above, the average amount of transactions seems to be higher in the same country for cards with fraudulent transactions.

The coefficient for the binary variable *Foreign Txn*, foreign currency transaction or not, was positive and significant. This shows that foreign currency transactions are more likely to be fraudulent. The exponentiated coefficient of 79.575 indicates that that a transaction being foreign multiplies the odds of it being fraudulent by nearly 80 times. The possible explanation is that the perpetrators operate globally, hoard credit card details of people around the globe and carry out transactions across countries.

Lastly, the binary variable *Ecommerce*, whether a credit card transaction is an electronic commerce transaction, had a positive and significant coefficient. This again indicates that transactions over the Internet have higher probability of being fraudulent. Perpetrators hoarding credit card details seem to prefer the anonymity of the Internet to commit fraud, as they do not have to be physically present before a merchant to carry out a transaction.

In Table 6 we present the summarized model estimation results. Our analysis showed that higher average spending per transaction over last month has a higher probability of fraud, while

higher spending over 3 months has lower probability of fraud. Higher average spending and higher number of transactions with the same merchant type over last month indicate higher chance of fraudulent behavior. However, higher average spending with the same merchant type over last 3 months indicates lower probability of fraudulent behavior. A higher number of transactions on the same day indicate fraudulent behavior, but a higher number of transactions with the same merchant same day has lower probability of fraudulent behavior. We also found that a higher number of transactions in the same currency over a month has higher probability of fraud. Further, higher average spending in a country over last month indicated higher probability of fraud, but higher number of transactions in the same country in last month indicated lower probability of fraud. Lastly, foreign currency transactions and Internet transactions had higher probability of fraud.

In Table 7, we present the classification table for estimated probability of fraud greater than 0.50. In other words, for any new transaction the model classified it to be fraudulent if the estimated probability of fraud exceeded 50%. The three common measures of performance of a model are sensitivity, specificity, and total percentage correct classification. Sensitivity of a model measures the percentage correct classification of fraudulent transactions as fraudulent and specificity measures the percentage correct classification of legitimate transactions as legitimate. The overall correct classification measures the percentage of correct classification out of the total number of transactions. The classification table shows that the model predicted 1838 out 2420 fraudulent transactions as fraudulent, i.e. the sensitivity of the model was 75.95%. Also, the model correctly predicted 12,302 out of 12,679 legitimate transactions as legitimate, i.e. the specificity of the model was 97.03%. The total percentage of correct classification for the logit model was 93.65%, i.e. 12,302 legitimate and 1838 fraudulent transactions out of the total of 15,099 transactions.

Table 6
Simplified results.

Independent variable	Consumer behavior	Direction of causality	Higher probability of
Txn amount over month	Higher average spending over last month	+	Fraud
Average over 3 months	Higher average spending over last 3 months	–	Legitimate
Amount merchant type over month	Higher average spending with the same merchant type over last month	+	Fraud
Number merchant type over month	Another transaction with the same merchant type over last month	+	Fraud
Amount merchant type over 3 months	Higher average spending with the same merchant type over last 3 months	–	Legitimate
Number same day	Another transaction same day	+	Fraud
Number same merchant	Another transaction with the same merchant same day	–	Legitimate
Number currency type over month	Another transaction in the same currency over a month	+	Fraud
Amount same country over month	Higher average spending in the same country over last month	+	Fraud
Number same country over month	Another transaction in the same country over last month	–	Legitimate
Foreign Txn	Foreign currency transaction	+	Fraud
Ecommerce	An Internet transaction	+	Fraud

Table 7
Classification table for the logit model.

	Predicted type		
	Legitimate	Fraudulent	Total
Observed type			
Legitimate	12,302	377	12,679
Fraudulent	582	1838	2420
Total	12,884	2215	15,099

Predicted type was fraud for estimated probability of fraud exceeding 0.5.

6. Discussion

In this study, we estimated the logit model using primary and derived attributes. We created derived attributes by aggregating values of transactions over different time periods. We found transaction aggregation to be a good strategy for fraud detection, as the model with derived attributes performed well in classifying transactions. The classification table showed that the logit model was adequate with respect to the percentage correct classification. However, overall percentage correct classification is not a very good measure of classification in case of studies like fraud where samples of legitimate cases far outnumber the fraud cases.

Detecting frauds in large datasets with small percentage of fraud cases, even with the best of algorithms, results in too many false positives, legitimate transactions identified as fraudulent (Krivko, 2010). The problem of high number of false positives is explained in (Levitt & Dubner, 2009) through an example of a detection method that is 99% accurate, i.e. this method correctly detects 99% of legitimate and fraudulent transactions. For example, in a dataset with 500 fraudulent and 50 million legitimate transactions, this detection method correctly identifies 495 fraudulent transactions, but incorrectly marks 500,000 legitimate transactions as fraudulent. In other words, for every single correctly identified fraudulent transaction, a supposedly accurate detection method with 99% accuracy incorrectly identifies more than 1000 legitimate transactions as fraudulent. This entails a huge cost in terms of manual follow up of all transactions identified as fraudulent.

Hence, sensitivity and specificity measures are used along with overall correct classification. The model can be made more sensitive by decreasing the threshold of classification from 0.50, but this will lower the specificity of the model. Lower specificity of the model means that more legitimate transactions will get flagged as fraudulent triggering potential loss of sales and unhappy customers. However, models with high sensitivity may find use with risk-averse merchants, who would prefer to catch more fraudulent transactions even at the expense of losing some potential sales by stopping legitimate transactions. Since, online merchants have different fraud loss risk tolerances and order rejection rates (CyberSource, 2008), logit models with higher sensitivity and therefore higher order rejection rates, may suit merchants selling high cost goods with low margins as the model is likely to err on the side of rejecting more orders to avoid expensive fraud losses. However, it is important to note that detection of fraud in large datasets requires judicious data partitioning and creation of effective derived attributes. We discuss these challenges in the following paragraphs.

Large financial institutions handle billions of credit card transactions and therefore it is important to partition their datasets to look for fraudulent transactions where likelihood of finding them is high, as it would be impossible to examine all transactions. For example, an online fraud report claims that foreign transactions in the US are four times more likely to be fraudulent than domestic transactions (CyberSource, 2009). Similarly, a select few categories of products, such as consumer electronics and jewelry, are more

likely to be targeted by perpetrators because they are easy to resell them online, for example on eBay (CyberSource, 2009). Hence, merchants need to continually observe customers' purchase behavior to focus on transactions where likelihood of fraud is high. In this study, we partitioned dataset A with all transactions to include only certain selected transaction types thereby reducing the dataset by half. Judicious data partitioning may not only make model estimation less challenging, but also improve the effectiveness of models.

Further, we would like to stress the importance of choice of derived attributes and their length of aggregation periods. Credit card transaction datasets have a number of categorical and numerical attributes. Hence the choice of attributes becomes very important in the efficacy of statistical models and this selection may change with change in fraudulent behavior. Also, the length of aggregation periods for these derived attributes is important.

We found that derived attributes aggregated over shorter period of time, a month, had positive coefficients indicating increase in probability of fraud with a unit increase in the coefficient, but for longer periods of aggregation, three months, coefficients were negative indicating decrease in probability of fraud. The reason for this could be that perpetrators usually try to use a stolen/ counterfeit card as much as possible during a short period before fraud is found and card is withdrawn. Hence, we should see abnormality in purchase behavior when we aggregate transactions for shorter period of time and we expect the abnormality to get softened in derived attributes with longer periods of aggregation. However, aggregation over too short a period may not capture sufficient information to raise alarm, for abnormality in spending could be by a genuine person (Whitrow et al., 2009). Hence, modelers have to be judicious about time periods while aggregating transactions for creation of derived attributes.

Lastly, one of the limitations of the study is the non-availability of time stamp of credit card transactions. We believe that availability of time stamp data for credit card transactions would have surely made the derived attributes much more discriminating in identification of fraudulent transactions. Due to lack of time stamp data we could not tell the order of transactions on any given day, so derived attributes had the same value for all transactions on any given day. For example, amount of transaction in a week prior to a transaction was same for all transactions in a given day as we could not tell the sequence of transactions due to lack of time stamp of transactions. We believe, with the availability of time stamp data, discriminatory power of models would have become better.

7. Conclusions

This study demonstrated the usefulness of creating derived attributes and judicious data partitioning for fraud detection. Practitioners and researchers may employ the idea of transaction aggregation in creation of suitable derived attributes. The key to creation of effective derived attributes is the choice of primary attributes and the length of aggregation periods of transactions. The choice of derived attributes may depend on the changes in perpetrators' fraudulent behavior over time [5] and future research may investigate this issue. Similarly, aggregation periods can also be an interesting area of future research and researchers may employ different aggregation periods for different attributes.

This study also highlights the importance of dataset partitioning to focus on transactions where frauds are more likely to occur. This is critical because it is impossible to verify all transactions, given the constraints of cost and time. Credit card transaction datasets are not only large but also have lopsided class sizes, with legitimate transactions far outnumbering fraudulent transactions. Hence, merchants have to focus on transaction types, product types, and/ or merchant types, where frauds are more likely to occur.

References

Artis, M., Ayuso, M., & Guillen, M. (2002). Detection of automobile insurance fraud with discrete choice models and misclassified claims. *The Journal of Risk and Insurance, 69*(3), 325–340.

Blunt, G. & Hand, D. J. (2000). The UK Credit Card Market. Technical Report, Department of mathematics, Imperial College, London.

Bollinger, C. R., & David, M. H. (1997). Modeling discrete choice with response error: Food stamp participation. *Journal of the American Statistical Association, 92*, 827–835.

Bolton, R. J., & Hand, D. J. (2001). Unsupervised profiling methods for fraud detection. In *Conference on credit scoring and credit control*, Edinburgh.

Bolton, R. J., & Hand, D. J. (2002). Statistical fraud detection: A review. *Statistical Science, 17*(3), 235–249.

Bose, R. (2006). Intelligent technologies for managing fraud and identity theft. In *Proceedings of the third international conference on information technology*, New Generations.

Breiman, L., & Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). *Classification and regression trees*. CA, Wadsworth, Belmont.

Caudill, S. B., Ayuso, M., & Guillen, M. (2005). Fraud detection using a multinomial logit model with missing information. *The Journal of Risk and Insurance, 72*(4), 539–550.

Chan, P. K., Fan, W., Prodromidis, A. L., & Stolfo, S. J. (1999). Distributed data mining in credit card fraud detection. Data Mining (November/December), 67–74.

Chen, R. C., Chen, T. S., & Lin, C. C. (2006). A new binary support vector system for increasing detection rate of credit card fraud. *International Journal of Pattern Recognition, 20*(2), 227–239.

CyberSource. (2008). Online Fraud Report: Online Payment, Fraud Trends, Merchant Practices, and Bench Marks, Retrieved January 8, from http://www.cybersource.com.

CyberSource. (2009). Online Fraud Report: Online Payment, Fraud Trends, Merchant Practices, and Bench Marks, Retrieved January 8, from http://www.cybersource.com.

Duman, E., & Ozcelik, M. H. (2011). Detecting credit card fraud by genetic algorithm and scatter search. *Expert Systems with Applications, 38*, 13057–13063.

Everett, C. (2009). Credit card fraud funds terrorism. Computer Fraud and Security, 1 May.

Hand, D. J. (1981). *Discrimination and classification*. Chichester: Wiley.

Hand, D. J. (1997). *Construction and assessment of classification rules*. Chichester: Wiley.

Hand, D. J., & Henley, W. E. (1997). Statistical classification methods in consumer credit scoring: A review. *Journal of Royal Statistics Society, 160*, 523–541.

Hausman, J. A., Abrevaya, J., & Scott-Morton, F. M. (1998). Misclassification of a dependent variable in a discrete-response setting. *Journal of Econometrics, 87*, 239–269.

Jin, Y., Rejesus, R. M., & Little, B. B. (2005). Binary choice models for rare events data: A crop insurance fraud application. *Applied Economics, 37*(7), 841–848.

Kou, Y., Chang-Tien, L., Sirwongwattana, S. & Huang, Y. P. (2004). Survey of fraud detection techniques. In *IEEE international conference on networking, sensing and control* (pp. 49–754).

Krivko, M. (2010). A hybrid model for plastic card fraud detection systems. *Expert Systems with Applications, 37*, 6070–6076.

Leggatt, H. (2008). CyberSource. Online Fraud to Reach $4 billion. BizReport, 16 December.

Leonard, K. J. (1993). Detecting credit card fraud using expert systems. *Computers and Industrial Engineering, 25*, 103–106.

Levitt, S. D., & Dubner, S. J. (2009). *Super freakonomics*. New York: HarperCollins.

Magder, L. S., & Hughes, J. P. (1997). Logistic regression when the outcome is measured with uncertainty. *American Journal of Epidemiology, 146*, 195–203.

McLachlan, G. J. (1992). *Discriminant analysis and statistical pattern recognition*. New York: Wiley.

Paasch, C. (2007). Credit Card Fraud Detection Using Artificial Neural Networks Tuned by Genetic Algorithms. Doctoral Dissertation, Hong Kong University of Science and Technology (HKUST), Hong Kong, "Unpublished Dissertation".

Phua, C., Lee, V., Smith, K., & Gayler, R. (2005). *A comprehensive survey of data mining-based fraud detection research*. Clayton School of information Technology: Monash University.

Poterba, J. M., & Summers, L. H. (1995). Unemployment benefits and labour market transitions: A multinomial logit model with errors in classification. *Review of Economics and Statistics, 77*, 207–216.

Quah, J. T. S., & Sriganesh, M. (2008). Real-time credit card fraud detection using computational intelligence. *Expert Systems with Applications, 35*, 1721–1732.

Quinlan, J. R. (1993). C4.5: Programs for machine Learning. San Mateo, CA, Morgan Kauffman.

Ripley, B. D. (1996). *Pattern recognition and neural networks*. Cambridge University Press.

Webb, A. R. (1999). *Statistical pattern recognition*. London: Arnold.

Wheeler, R., & Aitken, S. (2000). Multiple algorithms for fraud detection. *Knowledge Based Systems, 99*, 93–99.

Whitrow, C., Hand, D. J., Juszczak, P., Weston, D., & Adams, N. M. (2009). Transaction aggregation as a strategy for credit card fraud detection. *Data Mining and Knowledge Discovery, 18*(1), 30–55.

Williams, K. (2007). *The evolution of credit card fraud: Staying ahead of the curve*. eFunds Corporation.