Global Journal of Computer Science and Technology: C
Software & Data Engineering
Volume 16 Issue 4 Version 1.0 Year 2016
Type: Double Blind Peer Reviewed International Research Journal
Publisher: Global Journals Inc. (USA)
Online ISSN: 0975-4172 & Print ISSN: 0975-4350

Anti-Fraud Schema System for Identification and Prevention of Fraud Behaviors in E-Commerce Services

By Qinghong Yang, Wei Xing, Xiangquan Hu, & Yan Quan Liu

Southern Connecticut State University

Abstract - This study aims to determine the best practices and provide a model of the technical solutions that can effectively and systematically limit fraudulent transactions of online orders in e-commerce services, using the methods of analytical mining and case studies. Based on a process of fraud prevention and detection performed in the e-business Dangdang, Inc., a leading online retailer in China, twelve identifying features of fraudulent order data were extracted and compiled into a feature matrix. Logistic regression with this matrix was then used to build a model to judge if an order was fraudulent. The model was tested using various order data with machine learning techniques to meet the requirements of being effective, correct, adaptive, and persistent. Then an online detection and prevention schema was established and the hypothesis of so-called Behavior Pattern Change Assumption (BPCA) was proven.

Keywords: *e-commerce services, fraud behavior, determination, fraud prevention, case studies, logistic regression, machine learning.*

GJCST-C Classification : *K.4.4, H.2.1*

Strictly as per the compliance and regulations of:

Anti-Fraud Schema System for Identification and Prevention of Fraud Behaviors in E-Commerce Services

Qinghong Yang [α], Wei Xing [σ] , Xiangquan Hu, [ρ] , & Yan Quan Liu [ω]

Abstract- This study aims to determine the best practices and provide a model of the technical solutions that can effectively and systematically limit fraudulent transactions of online orders in e-commerce services, using the methods of analytical mining and case studies. Based on a process of fraud prevention and detection performed in the e-business Dangdang, Inc., a leading online retailer in China, twelve identifying features of fraudulent order data were extracted and compiled into a feature matrix. Logistic regression with this matrix was then used to build a model to judge if an order was fraudulent. The model was tested using various order data with machine learning techniques to meet the requirements of being effective, correct, adaptive, and persistent. Then an online detection and prevention schema was established and the hypothesis of so-called Behavior Pattern Change Assumption (BPCA) was proven. The results show the model can detect 94% of fraudulent orders. The Anti-fraud Schema System established for Dangdang is shown to be the best model for the determination and prevention of fraudulent behaviors in the e-commerce services.

Keywords: e-commerce services, fraud behavior, determination, fraud prevention, case studies, logistic regression, machine learning.

I. Introduction

Electronic commerce has enjoyed rapid growth in recent years [1], as more and more people have accepted online shopping. However, along with the growing number of transactions, there are a growing number of fraud activities. The temptation of economic gain and the difficulty of internet supervision have led to a great number of online fraud activities. Hackers can steal online accounts and use these accounts in criminal activities [2]. Prevention of fraud activities in order to provide a safe online shopping experience is a challenge for electronic commerce [3]. EBay is the leading e-commerce company around the globe, and every day thousands of customers trade through eBay. Therefore, eBay has hired experts from the National Aeronautics and Space Administration (NASA) of the U.S. to develop an anti-fraud model to detect and prevent fraud activities.

Author α σ ρ: School of Software, Beihang University, Beijing 100029, China.
Author ω: Southern Connecticut State University, New Haven, CT 06515, USA. e-mail: liuy1@southernct.edu

E-commerce started late in China, and few resources have been devoted to the anti-fraud field, so systematic anti-fraud solutions are especially scarce. As a leading business-to-consumer e-business in China founded in 1999, China Dangdang, Inc. offers products mainly in the categories of books, audios, digital devices, and household merchandise. Dangdang made an initial public offering (IPO) on the New York Stock Exchange in November 2010and had over 9,000,000 active customers. Because of its main business is online, Dangdang shows great interest in solving Internet-related fraud problems, especially those involving online orders directly affecting its customers, as its key strategy is to grow its e-business,

Beginning with are iew of past studies relevant to the present research, this paper provides detailed process descriptions of an anti-fraud model's development (Section 3), and discussions of its implementation with a real data mining process (Section 4). The results were concluded significantly that proven the best practice of fraud determination and detection in real situation of e-commerce order transactions.

II. Literature Review

A review by Hogan [4] summarized research on fraud behavior over the past decade. Prior to that researchers mainly focused on fraud in the areas of accounting, auditing and finance activities [5].The growth of online transactions led to a growth of fraud activities, and the lack of supervision online made it easy to commit fraud [3]. The online market has some unique features that attract fraud activities, namely information asymmetry, online transactions and the uncertainty of traders' identities and commodities [6].

Lou & Wang's research reveals that though many methods can be used to prevent fraud activities, e-commerce should use a systematic method to solve the problems uniformly [7]. Account information is collected and the information used to summarize the patterns of fraud to adapt to different situations [8]. Latch (1999) pointed out that using classification algorithms such as ID3 and C4.5 to detect fraud patterns and identify accounts with suspicious activity and then allowing humans to make the final judgment could work well in detecting fraud [9].

Detecting fraud behavior and managing fraud risk require the design and application of a fraud-detection model. Eining et al. find that auditors can manage different levels of fraud risk better and make unanimous auditing decisions by using an expert system [10]. Green and Choi use a neural network to detect fraud behavior and achieve satisfactory results [11].

Ohlson found that the identifying features of fraud activity could be used to alert sellers to fraud activities during financial transactions [12]. Lenard and Alam used logistic regression in detecting fraud activities [13]. This method has also been employed in several researcheslater on [14][15].

Maranzato [1] researches how to detect fraud activities in an e-commerce system. He then uses logistic regression to detect and identify features of credit fraud, and he also points out that logistic regression depends greatly on the data quality [16].

This work focuses on how to detect and prevent fraud activities in online transactions, especially before fraudulent orders are completed in the environment of e-commerce.

III. Research Design

This work combines analytical mining and case study using the real data of Dangdang's sale transactions to identify the common patterns of detecting and preventing fraud activities from occurring within online orders.

a) Collecting and Processing Data

Real customer order records collected from Dangdan's transaction logs, including initially identified fraudulent orders by the company's customer services officers, were processed in the following five phases:

Phase one: collect order data from Dangdang (01/01/2014-07/09/2014) and statistically analyze it to identify the features that can distinguish fraudulent orders from legitimate ones. Purposely, we used this method to find key features of fraud activities to and build a feature matrix system from machine learning.

Phase two: use the order data collected from the same period of time (01/01/2014-07/09/2014) to develop and train a logistic regression model to predict orders that are unusual.

Phase three: test the order data collected from Dangdang (07/01/2014-08/31/2014) with this logistic regression model to reveal how well the model works on the condition that the fraud ratio of orders is high.

Phase four: test the order data collected from Dangdang (10/07/2014-10/28/2014) with this logistic regression model again to reveal how well the model works on the condition that the fraud ratio of orders is low.

Phase five: employ this model in a non line environment with real customer records to assess the usefulness of the model.

b) Making Behavior Pattern Change Assumption

Unusual orders are associated with customer behaviors differing from normal ones. As empirical evidence in daily sales has accumulated, the researchers are convinced with such a set of hypothetical rules that may direct the fraud discovering process, as we so call Behavior Pattern Change Assumption (BPCA).

Rule 1, for most of the e-commerce user accounts, customer behaviors are consistent with shipping address, receiver name, receiver phone, payment habits, and so on, remaining steady. This is called 'steady behavior.' Sudden changes of some or all of these attributes may indicate fraudulent behavior.

Rule 2, when an order is confirmed as a fraudulent order, all the orders whose receiver address, receiver IP and so on are same as this order are considered suspicious. This is because one hacker may steal multiple accounts and make multiple orders; however, the IP and address may stay the same. It's like one fraudulent order infects the IP or address. This is known as the 'suspicion infection'.

Rule 3, hackers won't add their own money to an account but will just deplete the balance in the account or do other things that won't benefit the account but will deplete all possible resources from the account. They want to maximize their profit. This is called ''maximum rob".

c) Underling Research Procedure

An outline of the research procedure for this study consists of defining the case, analyzing the data, until extracting, evaluating and implementing the outcomes shown in Figure 3.

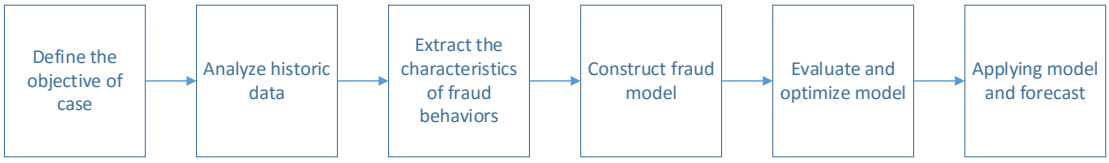

Figure 3: Research process of detecting fraudulent orders

Analyzing order data: The data from orders that have been marked as fraudulent by customer service are analyzed.

Feature extraction: Custom service and technology experts brainstorm to extract some features that may distinguish fraudulent orders from legitimate ones and statistically test them.

Model construction: Use the statistical analysis and apply an algorithm to the data and establish the norm of the algorithm. Use order data to create a logistic regression model.

Test and optimize the model: Use test order data to test and assess the performance of the model, then continue optimizing the model.

Model application: Use the fraudulent order detection model to judge online transaction orders then assess the performance and economy value of the model.

IV. Construction of a Fraudulent Orders' Detection Model

In an attempt to create effective technical solutions that could systematically limit fraudulent transactions of Internet orders, a fraudulent order detection model based on the Behavior Pattern Change Assumption (BPCA) was developed, consisting of the following process.

a) Determination of Fraudulent orders

Fraudulent orders occur when a hacker steals a customer's account and uses the balance in the account to purchase goods for him/her self. Fraudulent orders are confirmed when customers call customer service to complain. Customer service staff will also call the customer to check if the customer or hacker places an unusual order, which usually is the primary method to determine whether the order is a "regular" or "fraud" order.

b) Process of Analysis

The core idea of fraudulent order detection is to compare normal orders with fraudulent orders to find identifying features that distinguish them. These features can then be used to judge if an order is fraud or not with BPCA.

There are three steps to extract the features of fraudulent orders:

Step 1: Customer service staffs locate fraudulent orders because of customer complaints.

Step 2: Statistical analysis of commonly used information such as the IP address of the order, receiver name, receiver address and receiver phone number.

Step 3: Compare normal orders with fraudulent orders to identify features that distinguish fraudulent orders from normal ones.

c) Analysis of Source Data

Labelled order data are provided by Dangdang customer service, and analyzing these data can verify BPCA at some level. When a hacker places a fraudulent order, the receiver name, receiver number and receiver address are different from the normally used information. Because hackers don't want to use real addresses, they may use some generic rough ones ending with 'county', 'block', 'corner' or 'street'. Six features that can distinguish fraudulent orders from normal ones therefore were identified.

We analyzed the six features using real data from Dangdang to determine their effectiveness in identifying fraudulent orders. The source data were from Dangdang's order data between January 1 to July 9 of 2014, in a total of 2075 fraudulent orders and 1513 stolen accounts.

Table 4.1: Items that can distinguish fraudulent orders

item ID	Definition	Frequency
rough_addr	Is the address rough?	71.1%
usually_city	Is the receiver city usually used?	42.5%
usually_tel	Is the receiver phone usually used?	35.5%
usually_name	Is the receiver name usually used?	34%
usually_email	Is the receiver email usually used?	32.6%
payment_ratio,0.05	Pay for the order using extra money instead of the money stored in the account, the ratio of extra money is more than 0.05	7.3%

Taking the feature rough_ addr as example, frequency in the table means that 71.1% of all orders that have rough_ addr are fraudulent orders. According to the results of these statistical tests, a basic idea of fraudulent orders emerges and this result will help in building a machine-learning model. These results also reveal some interesting facts. First is that rough_ addr is a very identifiable feature from which to detect fraudulent order. Usually_ city, usually_ tel, usually_ name, usually _email all show some potential to detect fraudulent orders. Payment_ ratio verifies that hackers just want to maximize their profit, which is the rule 3 of BPCA. These simple statistical results are useful at some level, to solve the fraudulent order detection problem and to further verify that BPCA, the machine learning algorithm is needed.

d) Fraudulent order Detection Model using Logistic Regression

Firstly, Choosing an appropriate algorithm. Logistic regression is a suitable algorithm for fraudulent order detection because it is not hard to apply, and the company's customer service staff can easily interpret the result.

Assuming some identification features as previously described. The identifying features and order data are input into the regression, and then the algorithm builds a model where each feature has a coefficient showing how much this feature can affect the result. The features with low coefficients (i.e. that don't significantly affect the results) are removed and the model is run again.

Formula 4-1 shows the result of the model. If the model returns a result of 1, then the order is fraud, less than that indicating otherwise.

$$Y=\begin{cases} 1 \text{ fraud order} \\ 0 \text{ normal order} \end{cases} \quad (4\text{-}1)$$

Using identification features as x = (x_1,x_2,…,x_p), logistic regression can be represented as Formula 4-2.

$$Pf = P(Y=1|x) = \pi(x) = 1/(1+e^{(-g(x))}) \quad (4\text{-}2)$$

$$\text{Where } g(x) = \beta\_0+\beta\_1 x\_1+\beta\_2 x\_2+\cdots+\beta\_p x\_p \ (1\le p\le n)$$

Using maximum likelihood estimation, the coefficients β0, β1, βp can be obtained. According to Formula 4-2, an interested party can calculate the possibility of one order being a fraudulent order. When using a binary classifier, it becomes necessary to pick a threshold; if the possibility is greater than the threshold, the order is a fraudulent order, otherwise it may be innocent. The value of the threshold can be anywhere between 0 and 1.However, if the threshold is too low, the model would be unstable and if the threshold is too high, the recall rate would not be ideal. This paper chooses 0.75 as the threshold.

Second, conducting characteristic statistics and extraction. The most critical process to build the fraud detection model is to select identification features. The statistical information in Table 4.1 shows fraudulent orders always having features such as different receiver name, receiver address, receiver city, and receiver telephone number, which can ascertain fraudulent orders made by hackers going directly to their own addresses.

Features 1-14 in Table 4.2 are deduced by the concept "steady behavior" of BPCA. Receivers related information changes mean that this order might be a fraudulent order. Features 15-19 deduced by the concept "suspicion infection" of BPCA mean that if some receiver's phone numbers or receiver addresses have been complained about before, new orders that have the same receiver address and receiver phone number have the possibility of being fraudulent orders. Feature 20 is based on the statistical results shown in Table 4.1, if the receiver address is rough, there is a high possibility that the order is fraudulent order. Features 21-23 are based on "maximum rob" of BPCA, meaning that the hackers want to make the most profit possible out of the stolen account.

Table 4.2: Identification features of the logistic regression model

Feature ID	Feature name（x）	Explanation
1	name_dubious_count	Complaint number of this receiver name.
2	name_cust_dubious_count	How many customer IDs are related to this receiver name?
3	tel_home_dubious_count	Complaint number of this receiver telephone number.
4	tel_home_cust_dubious_count	How many customer IDs are related to this receiver telephone number?
5	tel_mobile_dubious_count	Complaint number of this receiver mobile phone number.
6	tel_mobile_cust_dubious_count	How many customer IDs are related to this receiver mobile phone number?
7	orderip_dubious_count	Complaint number of this receiver IP address
8	orderip_cust_dubious_count	How many customer IDs are related to this receiver IP address?
9	addr_dubious_count	Complaint number of this receiver address.
10	addr_cust_dubious_count	How many customer IDs are related to this receiver address?
11	permid_dubious_count	Complaint number of this receiver permid*
12	permid_cust_dubious_count	How many customer IDs are related to this receiver permid*
13	email_dubious_count	Complaint number of this receiver email.
14	email_cust_dubious_count	How many customer IDs are related to this receiver email?

15	name_frequency_count	How many orders does the customer make using this receiver name in history?
16	tel_home_frequency_count	How many orders does the customer make using this receiver telephone number in history?
17	tel_mobile_frequency_count	How many orders does the customer make using this receiver mobile number in history?
18	city_frequency_count	How many orders does the customer make using this receiver city in history?
19	addr_frequency_count	How many orders does the customer make using this receiver address in history?
20	rough_address	Is this address rough?
21	whole_price	The total price of the order.
22	Payment	How much money should the receiver pay when they receive this package?
23	payment_ratio	How much should be paid apart from using the account balance?
24	Intercept	The constant in logistic regression model.

*permid is used to identify customers. Whether or not the customer is logged in, Dangdang will save a permid on the device being used to browse Dangdang.

These identification features are used to compose an identification features matrix, and then training data is used to train the matrix. Some of the identification features may not be effective in detecting fraudulent orders, and some ineffective features are eliminated during the training process.

V. Application and Results

The implementation of the logistic regression discussed above helps develop a fraudulent order detection model and test its effectiveness using the real order data of Dangdang as the experimental subject.

a) Preparation and Preprocessing of Data

Preprocessing of data is a key problem in machine learning, because in most cases data is incomplete, noisy and incompatible. The result of a machine-learning algorithm depends greatly on the quality of data. Data preprocessing includes: data cleaning, data integration, data conversion and data reduction [17].

Because of the volume of data, a sample of the total order data has been used with a ratio of fraudulent orders versus normal orders of from 1:5 to 1:9.

Continuous numbers were assigned to discrete sections. For example, the total money was divided into sections [0,10), [10,50), [50,100), and >100. Discretization can be used when the focus is only on relative value instead of absolute value. The discretization formula used in this paper is ln (x+1)/ln2. Discretization is useful to describe nonlinear relationships and solve the hidden flaws in data [11].

b) Process and Application of Model

Based on Formulae 4-1 and 4-2, R programming language was used to create a logistic regression model and then train the model to obtain coefficients.

$$g(x) = \beta\_0+\beta\_1 x\_1+\beta\_2 x\_2+\cdots+\beta\_p x\_p(1\leq p\leq 23) \quad (5\text{-}1)$$

After training, the coefficients, namely β0, β1, …βp can be figured out as shown in Table 5.1.

Table 5.1: Coefficients of first training

Index	X (Feature name)	B (coefficient)
1	email_dubious_count	6.990
2	name_cust_dubious_count	0.619
3	rough_address	0.532
4	orderip_cust_dubious_count	0.168
5	permid_cust_dubious_count	0.166
6	addr_dubious_count	0.075
7	tel_mobile_dubious_count	0.062
8	tel_home_cust_dubious_count	0.049
9	addr_cust_dubious_count	0.003
10	whole_price	0.000
11	name_frequency_count	0.000
12	email_cust_dubious_count	0.000
13	Payment	-0.000
14	addr_frequency_count	-0.004
15	tel_mobile_frequency_count	-0.006
16	tel_home_dubious_count	-0.023

17	payment_ratio	-0.030
18	tel_mobile_cust_dubious_count	-0.031
19	permid_dubious_count	-0.056
20	orderip_dubious_count	-0.097
21	city_frequency_count	-0.341
22	name_dubious_count	-0.462
23	tel_home_frequency_count	-0.464
24	Intercept	-5.161

c) Optimization and Second Training of Model

After first training, the coefficients of the first model were obtained. Then, based on the analysis of these results, 13 features were deleted and 1 new feature added. Features were deleted based on three rules:

Rule 1: If one feature's coefficient is 2 magnitudes lower than the biggest coefficient, it should be deleted.

Rule 2: If one feature's coefficient is not logical, it should be deleted.

Rule 3: If one feature is considered not logical after discussion with experts, it should be deleted. Features that have the pattern "*** _cust_dubious_count" all have low coefficients and after discussion it was determined that these features are not very logical, so they were deleted (see Table 5.2).

Table 5.2 : Features that were deleted and why

X (Feature name)	β(coefficient)	Reason for deletion
Payment	-0.000	r1
addr_cust_dubious_count	0.003	r3
email_cust_dubious_count	0	r3
name_dubious_count	-0.462	r2
name_cust_dubious_count	0.619	r3
orderip_cust_dubious_count	0.168	r3
permid_cust_dubious_count	0.166	r3
tel_home_dubious_count	-0.023	r1
tel_home_cust_dubious_count	0.049	r3
tel_mobile_cust_dubious_count	-0.031	r3
name_frequency_count	0	r1
tel_home_frequency_count	-0.464	r3
tel_mobile_frequency_count	-0.006	r1

A new feature, phone_address was added. Phone_address is a complex feature that can be calculated by this rule: if neither the receiver mobile number nor the address of one order have ever been used in this account, then phone_address is the number of total history orders of this account, unless phone_address is 0. This feature was added based on the reasoning that the more orders a user has bought, the lower the likelihood of them changing receiver address and mobile number at the same time. The final logistic model is shown in Table 5.3.

Table 5.3 : Final features and coefficients of logistic model

Sequence	X (Feature name)	β(coefficient)
1	city_frequency_count	-0.405
2	addr_dubious_count	0.305
3	email_dubious_count	2.680
4	orderip_dubious_count	0.561
5	phone_address	0.887
6	tel_mobile_dubious_count	0.993
7	whole_price	0.338
8	addr_frequency_count	-1.050
9	payment_ratio	-0.200
10	Intercept	-1.395
11	permid_dubious_count	0.605
12	rough_address	0.406

By analyzing the final model, BPCA is testified and the following conclusions are reached.

The coefficients of city_frequency_count and addr_frequency_count are negative, which means if the receiver city and receiver address of an order have been used in this account many times, the order is less suspicious.

The coefficients of addr_dubious_count, email_dubious_count,tel_mobile_dubious_count,permid_dubious_count and orderip_dubious_count are all positive.

This result confirms the hypothesis of "suspicion infection": the new orders that have the same receiver address, mobile number and IP as previous fraudulent orders are considered suspicious.

The coefficient of rough_address is positive, which means that an order with a rough address is suspicious because those committing fraud do not want to supply their address.

The coefficient of par_rate is negative, which means that if the customer should pay extra money over the value of their account balance, the order is less suspicious.

The coefficient of whole_price is positive, which means that fraudulent orders tend to be greater in total price. However, the absolute value of the coefficient is one of the lowest ones, so this tendency is not very important.

The coefficient of phone_address is positive, which means that if one account has made many orders in history and now uses both a new receiver mobile number and a new address to make a new order, then the new order is suspicious.

The results shown in table 5.3 verify the BPCA.

d) Test and Performance of Model

Using the values (x1, x2, x4….) of the features of one order as input, the model calculated a possibility (Pf as shown in Formula 5-2) of this order being a fraudulent order. If the possibility is greater than the threshold (0.75), then this order is a fraudulent order, otherwise legitimate.

$$Pf = P(Y=1|x) = \pi(x) = 1/(1+e^{(-g(x))}) \quad (5\text{-}2)$$

$$g(x) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p \quad (1 \le p \le 11)$$

The values of β0, β1, … βp are shown in Table 5.3. X1, x2…xp are provided by Dangdang's data system.

Four categories were defined: FF (judged as fraud, in fact is fraud), FC (judged as fraud, in fact is clear), CF (judged as clear, in fact is fraud), and CC (judged as clear, in fact is clear). Human check number is the number of orders that should be checked by customer service officers, recall rate is the rate of fraudulent orders that can be detected by the system, the calculation of these two values are shown below.

$$\text{Human check number} = FF + FC \quad (5\text{-}3)$$

$$\text{Recall rate} = FF/(FF+CF) \quad (5\text{-}4)$$

Two experiments were designed to test the effects of the fraud detection model in different conditions of time and situation.

Experiment 1 used Dangdang order data from 07/01/2014 to 08/31/2014 as the test data set. It detected 395 of a total of 417 fraudulent orders. The result is shown in Table 5.4 where it can be seen that the model works well in a situation where the rate of fraudulent orders is high.

Table 5.4 : Results of Experiment 1

Index	Value	Index	Value
Fraudulent order number	417	CF	22
Normal order number	346725	FC	896
Total order number	347142	CC	345829
Threshold	0.75	Recall	0.95
FF	395	Human check number	1291

Experiment 2 used Dangdang order data from 10/07/2014 to 10/28/2014 as the test data set. The model detected 48 of a total of 51 fraudulent orders. The result is shown in Table 5.5. Three fraudulent orders were missed. One of them was the first order of a new account, and the other two were a situation in which one customer used another customer's gift card, but the account using the gift card was not stolen.

Table 5.5 : Results of Experiment 2

Index	Value	Index	Value
Fraudulent order number	51	CF	3
Normal order number	270414	FC	1443
Total order number	270465	CC	268971
Threshold	0.75	Recall	0.94
FF	48	Human check number	1491

The fraudulent order rate in Experiment 2 is 0.000189, which was less than the rate 0.001201 of Experiment 1. It can be seen that the recall rate and human check number are close to that in Experiment 1.

e) Performance of the Model

The changes in numbers of fraudulent orders and amount of money stolen were analyzed to show the usefulness of the fraudulent order detection system. The anti-fraud system using this fraud detection model started to run on 06/24/2014. Figure 5.1 shows the change in fraudulent order numbers from 01/2014 to 01/2014. Figure 5.2 shows the change in amount of money being stolen from 01/2014 to 01/2014. Based on the information in Figures 5.1 and 5.2, it appears that after the system was implemented in 06/2014, the fraudulent order problem was controlled. Such a fraud detection system was not in place in Dangdang before, as shown in figure 5.1. Prior to April 2014 there were not many fraudulent orders found, because Dangdang had no way to control the situation. April 2014 could be a bench mark in China. That year a number of accounts were terribly leaked in China. Dangdang, as one of the biggest ecommerce entities, were attacked by hackers who used the leaked accounts

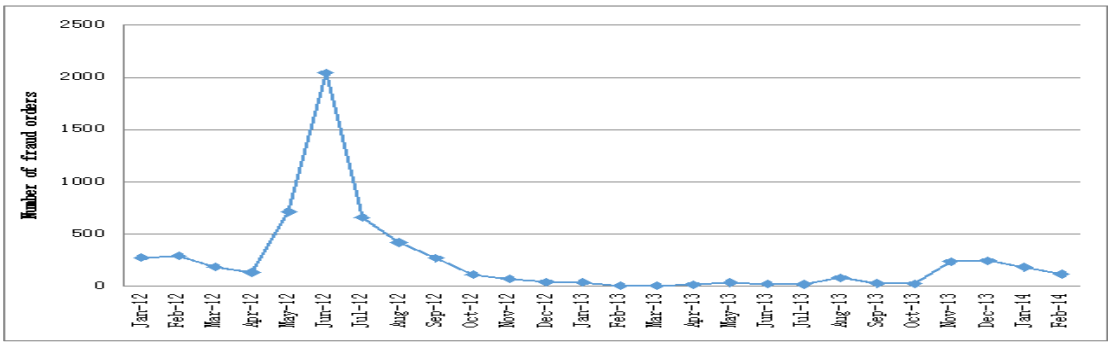

Figure 5.1 : Trend of fraudulent order number per month

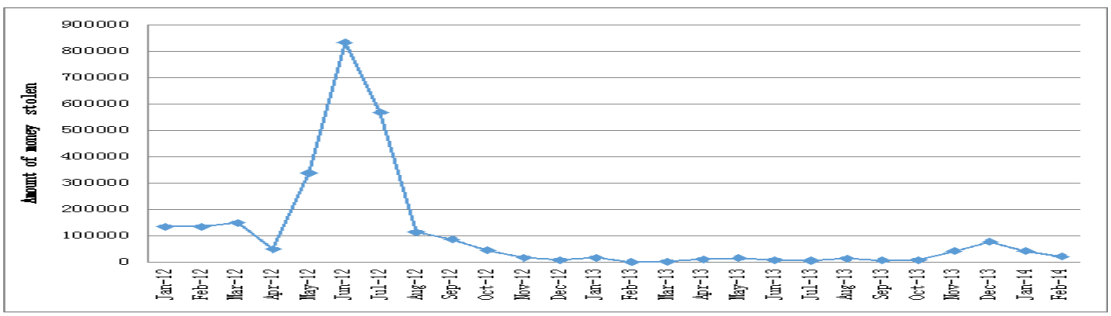

Figure 5.2 : Amount of money being stolen per month

94% of all fraudulent orders can be detected by this system. After the anti-fraud detection system was implemented, fewer complaints were received. Combining this system with human rechecking, both the number of fraudulent orders and the amount of money stolen were reduced.

VI. Conclusions

We found that trading fraud in online ecommerce can be detected and prevented. Order fraud is a key problem of trading online and is very harmful to companies. By focusing on the determination of these online order frauds, the real situation and data of Dangdang was used with these two steps to follow.

First, statistical analysis was carried out to determine the basic differences between fraudulent orders and normal ones, and then a logistic regression model was created. Finally, different experiments were designed to test the validity and effectiveness of the Fraudulent orders' Detection Model using twelve identifying features, which is the best model approved by the company in identification and prevention of fraud behaviors in e-commerce services.

a) *Pattern of Detecting and Preventing Online Seller and Customer Fraudulent orders*

First, statistical analysis of order data was carried out to provide a basic idea of the characteristics of fraudulent orders. Then, the features that can help in distinguishing fraudulent orders from normal ones were extracted and used to format a feature matrix. The feature matrix and logistic regression algorithm were used to build a fraudulent order detection model and carry out optimizations. Finally, the model's effectiveness was tested, and the model was used to detect real time orders and keep track of the performance of the model.

There would be no way to determine fraudulent orders without the implementation of this new developed model, which has been allowing enabled the company's customer service staff to successfully catch and free zesuspicious accounts. About 94% of fraudulent orders were detected in Dangdang in the past year with the model. It helped the company reduce fraudulent orders and therefore could be instructive to and implemented by similar electronic commerce entities.

b) *Economical Significance of Detecting and Preventing Fraudulent order*

Reducing the number of fraudulent orders can benefit both customers and companies. First, fewer fraudulent orders means fewer customers losing money and more customers enjoying their shopping experience. Second, fewer fraudulent orders means that companies will receive fewer complaints and customer satisfaction will be higher. The public image of companies is improved with fewer fraudulent orders.

VII. Future Work

This report describes innovative research on the role of features of online orders and accounts in monitoring online transaction activities. This work resulted in a model that can detect patterns of fraud activity in online transactions, and judges the likelihood of each transaction being fraudulent. When the likelihood of a transaction being a fraud is high, human checks are still required to make a final judgment. Therefore, human resources are needed in identifying fraud. There are a variety of fraud activities, so different detection processes are required. Therefore, the process and model of this work can in future be adjusted to be used in more situations.

Acknowledgements This research was supported by Dangdang (http://www.dangdang.com). Thanks to Fu Qiang, VP of Technology, Ju Qi, Technical director, and other staff of Dangdang for their support and help. Thanks to the technical help of Michael Wagner of Yale University Social Sciences Information Center. Thanks to Hu Saiquan of Yale University Economics and Management School, and Sam of Yale Computer Science School for discussion of research methods. Thanks to Ms. Arlene and BiBielefield, as well as Mr. Keven and Quan Wang for corrections to the paper.

References Références Referencias

1. Jans, M., N. Lybaert, and K. Vanhoof. 2009. A framework for internal fraud risk reduction at IT integrating business processes: The IFR framework. The International Journal of Digital Accounting Research 9: 1–29.
2. Kim, T.K., Lim, Y.J. & Nah, J.H. (2013) Analysis on fraud detection for internet service. International Journal of Security and Its Applications, 7 (6): 275-284.
3. Shim, S. & Lee, B. (2010) An economic model of optimal fraud control and the aftermarket for security services in online marketplaces. Electronic Commerce Research and Applications, 9(5): 435-445.
4. Hogan, C. E., Rezaee, Z., Riley, R.A. & Velury. U.K. (2008) Financial statement fraud: Insights from the academic literature. Auditing: A Journal of Practice & Theory, 27(2): 231–252.
5. Trompeter, G.M., Carpenter, T.D., Desai, N., Jones, K.L. & Riley Jr., R.J. (2013) A synthesis of fraud-related research. Auditing: A Journal of Practice & Theory, 32 (S1): 287-321.
6. Klein, B., & Leffler, K. B. (1981) The role of market forces in assuring contractual performance. The Journal of Political Economy, 89 (4): 615–641.
7. Lou, Y.I. & Wang, M.L.(2009) Fraud risk factor of the fraud triangle assessing the likelihood of fraudulent financial reporting. Journal of Business & Economics Research, 7(2): 61–78.
8. Michael & Adler (1971) pointed out that the sole concern of fraud risk factor studies was the consideration of fraudulent behavior or how to detect or to deter fraud.
9. Lach, J. (1999) Data mining digs in. American Demographics, 38-45.
10. [JW1] Eining, M.M., Jones, D.R., & Loebbecke, J.K. (1997) Reliance on decision aids: an examination of auditors' assessment of management fraud. Auditing: A Journal of Practice and Theory, 16 (2): 1-19.
11. Green, B.P. & Choi, J.H. (1997) Assessing the risk of management fraud through neural network

technology. Auditing: A Journal of Practice and Theory, 16 (1): 14–28.
12. Ohlson, J.A. (1980) Financial ratios and the probabilistic prediction of bankruptcy. Journal of Accounting Research, 18: 109–131.
13. Lenard, M.J. & Alam, P. (2009) An historical perspective on fraud detection: From bankruptcy models to most effective indicators of fraud in recent incidents. Journal of Forensic & Investigative Accounting, 1(1): 1-5.
14. Zhang, H., Lin, Z., & Hu, X. The effectiveness of the escrow model: an experimental framework for dynamic online environments. Journal of Organizational Computing and Electronic Commerce, 17 (2): 119–143.
15. Hosmer, D.W., Lemeshow, S. & Sturdivant, R.X. (2013) Applied Logistic Regression. Wiley & Sons, New York.
16. Maranzato, R., Pereira, A., do Lago, A.P. & Neubert, M. (2010) Fraud detection in reputation systems in e-markets using logistic regression. Proceedings of the 2010 ACM Symposium on Applied Computing (SAC), Sierre. Switzerland.
17. Lek, M., Anandarajah, B., Cerpa, N. & Jamieson, R. (2001) Data mining prototype for detecting e-commerce fraud. The 9th European Conference on Information Systems, Bled, Slovenia.