A Novel Protocol for Detection and Prevention of Fraud in Online Through Proactive Model

Krishnaveni Ampolu [#1], P.Srinivasu Varma [*2]
[#1]M.Tech Scholar, [*2]Assistant Professor
Department of Computer Science & Engineering,
Avanthi's St.Theressa Institute of Engineering & Technology,
Garividi, Vizianagaram Dist, AP, India.

Abstract

In the Past 4 Years, e-commerce is growing faster than predicted as it is up over 400% compared in the past. As customers have the ease to buy things without spending much time there are also some criminals who try to do fraud and get profit in illegal ways. As people are enjoying the advantages from online trading, traitors are also taking advantages to accomplish betrayal activities against candid parties to obtain dishonest profit. In this paper in order to detect and prevent such illegal and betrayal activities online probit fraud-detection moderation systems are commonly applied in practice. Machine learned models which are learned online are capable to catch deceptive more proficiently and quickly than human-tuned rule-based systems. In this paper, we show that this model can probably distinguish more deceptions and extensively decrease customer complaints which are based on a real-world online auction fraud detection data compared to several baseline models and the human-tuned rule-based system.

Keywords

World Wide Web (WWW), Online Shopping, Fraud Detection, Online Probit Model, Online Auction, Machine Learned Models, Betrayal, Moderation.

1. Introduction

Nowadays various websites are allowing Internet users to purchase and trade products which profits everyone in terms of effortlessness and effectiveness [1], [2]. People who shop are enjoying the advantages from online trading; at the same time traitors are also taking advantages to accomplish deceptive activities against candid parties to obtain dishonest profit. Since the emergence of World Wide Web (WWW) as in [3], electronic commerce, which is commonly called as e-commerce as in [4] become more and more popular in recent years. No often do we now think of taking a stroll through the super market before buying a mobile handset, but a healthy online research which in some cases is consequently followed by an online purchase. The scenario is not limited to mobiles alone; it even covers a wide range of products like home appliances, consumer electronic goods, books, apparels, travelling packages etc. and even the electronic content itself. With e-Commerce as in [4] then, you can buy almost anything you wish for without actually touching the product physically and inquiring the salesman for a number of times before placing the final order. In existing online shopping business model sellers as in [5] sell their products or

services at preset price, where buyers can choose what product best suites them which is of good deal.

Online auction however is a different business model where the items are sold through price bidding done by various users. Usually bidding have starting price and expiration time prescribed for every product. Potential buyers in auction bid against each other, and the winner is the one who bids the item for highest price among all buyers. To provide some assurance against fraud and to give confidence to online auction services as in [6] E-commerce sites provide insurance to victims for those who loss up to a certain amount. Online auction services and ecommerce sites adopt following approaches to control and prevent fraud occurred in online shopping.

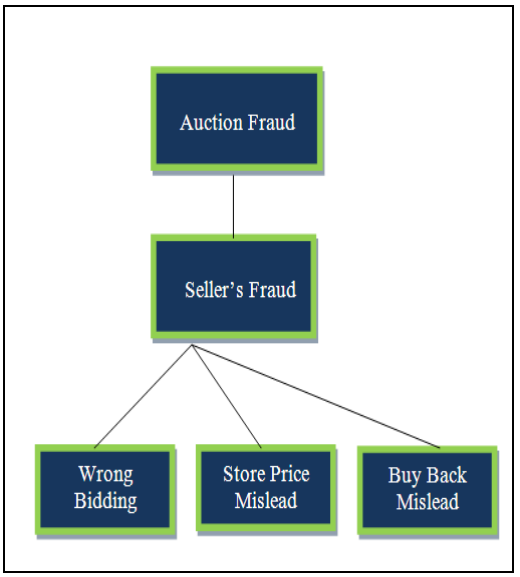

Figure. 1. Represents Online Auction Fraud

For purchasing a certain item from the online auction website they are to be validated with e-mail, SMS, or phone call verifications as in [7]. In this paper, we study the application of a proactive moderation system as in [8] for fraud detection, where hundreds and thousands of new auction cases are created every day. Due to the limited expert resources only 20%-40% of cases can be reviewed and labeled but not all. Therefore, it is necessary to develop pre-screening moderation system as in [8] that only directs suspicious/illegal cases for expert review and passes the rest as clean cases. Human experts are also willing to test and see the results of online feature selection to monitor the effectiveness and stability as in [9] of the current set of features, so that they can understand the pattern of frauds done by fraudulent sellers and further add or remove some features as shown in Figure. 1.

In this paper we study the problem of building online modeling system for the auction fraud detection moderation system as in [10]. We propose a Bayesian online fraud detection model framework for the binary response. We apply the stochastic search variable selection (SSVS) as in [11], a well known technique to handle statistical literature, to handle the dynamic evolution of the feature importance in a principled way. Similar to as in **[12],** we consider the expert knowledge to bound the rule-based coefficients [13] to be positive.

2. Related Work

The main important issue in the online is fraud/fault that is caused in online shopping. There are many articles on websites which teach people how to avoid online auction fraud [14] categorizes auction fraud into several types and proposes strategies to fight them. Some Reputation systems are used extensively by web-sites to detect auction frauds, although many of them use naive approaches. [15] Summarized several key properties of a good reputation system and also the challenges for the modern reputation systems to elicit user feedback. Other representative work connecting reputation systems with online auction fraud detection include [16, 17, 18], where the last work [18] introduced a Markov random field model with a belief propagation algorithm for the user reputation.

In this paper we treat the fraud detection problem as a binary classification problem which has two possibilities. The most frequently used models for binary classification include logistic regression [19], probit regression [20], support vector machine (SVM) [21] and decision trees [22]. Feature selection for regression models is often done

through introducing penalties on the coefficients. Typical penalties include ridge regression [23] (L2 penalty) and Lasso [24] (L1 penalty). Compared to ridge regression, Lasso shrinks the unnecessary coefficients to zero instead of small values, which provides both intuition and good performance. Stochastic search variable selection (SSVS) [11] uses "spike and slab" prior [25] so that the posterior of the coefficients have some probability being 0.

Another approach is to consider the variable selection problem as model selection, i.e. put priors on models (e.g. a Bernoulli prior on each coefficient being 0) and compute the marginal posterior probably of the model given data. People then either use Markov Chain Monte Carlo to sample models from the model space and apply Bayesian model averaging [26], or do a stochastic search in the model space to find the posterior mode [27]. Among non-linear models, tree models usually handles the non-linearity and variable selection simultaneously. Representative work includes decision trees [22], random forests [9], gradient boosting [28] and Bayesian additive regression trees (BART) [27].

3. Proposed Methodology

In this paper we have proposed following new methodologies as proposed techniques for Online Fraud Detection.

3.1 Online Equity Regression Methodology

Consider splitting of continuous time into many small equal size intervals as in [11]. For every interval we may have many expert labeled cases which indicate whether they are fraud or not. At time interval t suppose there are **nt** observations. Let us denote the i-th binary observation as y_{it}. If $y_{it} = 1$, the case is fraud, otherwise it is non-fraud. Let the feature set of case i at time t be X_{it} . The online fraud detection model as in [12] can be written as

$$P [y_{it} = 1 | x_{it} , \alpha t] = \Phi (x`_{it} , \alpha t) , \quad (1)$$

Where $\Phi(\cdot)$ is the cumulative distribution function of the standard normal distribution N (0, 1), and αt is the unknown regression coefficient vector at time t. By data augmentation the online fraud detection model can be expressed in a hierarchical form as follows:

For each observation i at time t assume a latent random variable z_{it} .The binary response y_{it} can be viewed as an indicator of whether zit > 0, i.e. $y_{it} = 1$ if and only if $z_{it} > 0$. If $z_{it} <= 0$, then $y_{it} = 0$. z_{it} can then be modeled by a linear regression as in **[19]** .

$$z_{it} \sim N (x'_{it}\alpha t, 1). \quad (2)$$

In a Bayesian modeling framework it is common practice to put a Gaussian prior on αt.

$$\alpha t \sim N (\mu_t , \Sigma_t), \quad (3)$$

Where μt and Σt are prior mean and prior covariance matrix respectively.

3.2 Online Feature Selection through SSVS Methodology

For regression problems with many features, proper shrinkage on the regression coefficients as in [21] is usually required to avoid over-fitting. For instance, the two common shrinkage methods are L2 penalty (ridge regression) and L1 penalty (Lasso) as in [13]. Experts often want to monitor the importance of rules so that if any adjustments are required they can modify it for effective use. By this experts as in [11] can add new rules or change rules. However, the fraudulent sellers change their behavioral pattern quickly: some rule-based features that does not help today might help a lot tomorrow. For that it is necessary to build an online feature selection framework and intuition as in [16]. At time t, let α_{jt} be the j-th element of the coefficient vector αt. Instead of putting a Gaussian prior on α_{jt} , the prior of α_{jt} now is as in [17].

$$\alpha jt \sim p_{0it}\, 1(\alpha jt = 0) + (1 - P_{0it})N(\mu_{it}, \sigma^2_{it}), \quad (4)$$

Where p_{ojt} is the prior probability of α_{jt} being exactly 0, and with prior probability $1 - p_{ojt}$, α_{jt} is drawn from a Gaussian distribution with mean μ_{jt} and variance σ_{jt}^2. Such prior is called the "spike and slab" as in [8] but how to embed it to online modeling has never been explored before.

3.3 Multiple Instance Learning Methodology

In the modern system the expert labeling procedure is in a bagged fashion as in [13] i.e. when a new labeling process starts, an expert picks the most suspicious seller in the queue and looks through all his/her cases posted in the current batch; the expert determines if any of the cases had been found to be fraud, then all the cases from this seller are labeled as fraud. In these types of scenarios they are to be handled by "multiple instance learning" as in [2]. Suppose for each seller i at time t there are K_{it} number of cases. For all the K_{it} cases the labels should be identical, hence can be denoted as y_{it} .For probit link function, through data augmentation denote the latent variable for the l-th case of seller i as z_{ilt} .The multiple instance learning model can be written as

$$y_{it} = 0 \text{ iff } z_{ilt} < 0, \forall l = 1, \cdots, K_{it} \quad (5)$$

Otherwise

$$y_{it} = 1, \text{ and } z_{ilt} \sim N(x'_{ilt}\alpha t, 1); \quad (6)$$

4. Implementation Modules

Implementation is the stage of the project when the theoretical design is turned out into a working system. Thus it can be considered to be the most critical stage in achieving a successful new system and in giving the user, confidence that the new system will work and be effective. The implementation stage involves careful planning, investigation of the existing system and it's constraints on implementation, designing of methods to achieve changeover and evaluation of changeover methods.

The proposed consists of totally 4 Modules:

1) Registration & Authentication
2) Login Module
3) Fraud Churn
4) User Complaint

1) Registration & Authentication Module

In this module we describe about user to be registered for accessing the site. The user may be a seller, user.When user comes first time for browsing, this module will create new profile registration to that particular user. After profile registration is successful with all valid field and user id and password, the user will be entered into the model based on his role either user or seller if he is giving a valid userid and password, if not login fails.

2) Login Module

In this module, after successful registration of all sellers and users, the administrator login into the model with his user id and password for performing various operations like view all sellers and giving authorization access for the newly registered sellers. If the authorized access is not given to sellers by admin,even they have userid and pass word, they can't able to enter into their individual accounts for creating new products and offers for the products. Hence if the authorization is given by admin to the seller then he/she can able to place new products and offers for online shopping available users who wish to do purchase in this site.

3) Fraud Identification Module

In this module, if one case is labeled as fraud by human experts that are maintained by administrator, it is very likely that the seller is not trustable and may be also selling other frauds; hence all the items submitted by the same seller are labeled

as fraud too. The fraudulent seller along with his/her cases will be removed from the website immediately once detected.

4) User Complaint Module

Buyers can file complaints to claim loss if they are recently deceived by fraudulent sellers. The administrator views the various type of complaints like Not Delivered, Product Mismatch, Poor Service and Product Damaged, and the percentage of various type complaints. The complaints values of a products increase some threshold value the administrator set the trustability of the product as Untrusted or banded. If the products set as banaded, the user cannot view the products in the website.

5. Experimental Results

We conduct our experiments on a real online auction fraud detection data set collected from a major Indian website. We consider the following online models:

• **ON-PROB** is the online probit regression model described in Section 4.1.

• **ON-SSVSB** is the online probit regression model with "spike and slab" prior on the coefficients, and the coefficients for the binary rule features are bounded to be positive (see Section 4.2).

• **ON-SSVSBMIL** is the online probit regression model with multiple instance learning and "spike and slab" prior on the coefficients. The coefficients for the binary rule features are also bounded to be positive (Section 2.3).

5.1 Evaluation Metric

In this paper we adopt an evaluation metric introduced in that directly reflects how many frauds a model can catch: the rate of missed complaints by admin, which is the portion of customer complaints that the model cannot capture as fraud. Note that in our application, the labeled data was not created through random sampling, but via a pre-screening moderation system using the expert-tuned coefficients (the data were created when only the expert model was deployed). This in fact introduces biases in the evaluation for the metrics which only use the labeled observations but ignore the unlabeled ones. This rate of missed complaints metric however covers both labeled and unlabeled data since customers do not know which cases are labeled, hence it is unbiased for evaluating the model performance as shown in Figure 2.

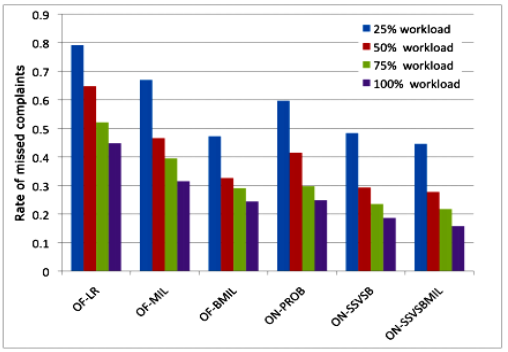

Figure 2: The rates of missed customer complaints for workload rates equal to 25%, 50%, 75% and 100% for all the offline models and online models with daily batches.

5.2 Model Performance

We ran all of the offline and online models on our real auction fraud detection data and show the rates of missed customer complaints given 100% workload rate. Note that for online models we tried δ (one key parameter to control how dynamically the model evolves) for different values (0.6, 0.7, 0.75, 0.8, 0.9, 0.95 and 0.99) and report the best in the table. For ω we also did similar tuning and found that $\omega = 0.9$ seems to be a good value for all models. From the table it is very obvious that the online models are generally better than the corresponding offline models (e.g. ON-PROB versus OF-LR, ON-SSVSBMIL versus OF-BMIL), because online models not only learn from the September training period but also update for every batch during the May 2013 test period as it is shown in Table 1.

δ	0.75	0.8	0.9	0.95	0.99
Rate of Missed Complaints	0.1463	0.1330	0.1641	0.1729	0.1973

Table.1.The rates of missed customer complaints for ON-SSVSBMIL (100% workload rate, batch size equal to 1/2 day and w = 0.9), with different values of δ.

6. Conclusion

In this paper, the designed model is almost used for most important online auction website; we build online models for the auction fraud detection and avoidance of fraud. We show that our proposed online proactive model framework is based on a real word online auction fraud detection data, which combines online feature selection, bounding coefficients from proficient knowledge and several instance learning and can extensively develop over baselines and the human-tuned model. This online modeling frame can be simply extended to several other applications like web spam detection, content optimization and so forth. . The adjustment of the selection bias in the online model training process is included to one direction and has proven to be very efficient for offline models.

7. References

[1] Agarwal D, Chen B, Elango P. Spatio-temporal models for estimating click-through rate. In Proceedings of the 18th international conference on World wide web ACM 2009; 21-30.

[2] Andrews S, Tsochantaridis I, Hofmann T. Support vector machines for multiple-instance learning. Advances In Neural Information Processing Systems 2003; 577-584.

[3] D. AGARWAL, B. CHEN, AND P. ELANGO. SPATIO-TEMPORAL MODELS FOR ESTIMATING CLICK-THROUGH RATE. IN PROCEEDINGS OF THE 18TH INTERNATIONAL CONFERENCE ON WORLD.

[4] D. Gregg and J. Scott. The role of reputation systems in reducing on-line auction fraud. International Journal of Electronic Commerce. 10(3):95-120, 2006.

[5] Federal Trade Commission. Internet auctions: A guide for buyers and sellers. http://www.ftc.gov/bcp/conline/pubs/online/auctions .htm, 2004.

[6] D. Chau and C. Faloutsos. Fraud detection in electronic auction. In European Web Mining Forum (EWMF 2005), page 87.

[7] A. Borodin and R. El-Yaniv. Online computation and competitive analysis, volume 53. Cambridge University
Press New York, 1998.

[8] H. Ishwaran and J.Rao. Spike and Slab variable selection: Frequentist and Bayesian strategies. The Annals of Statistics,33(2): /730-733 ,2005.

[9] L. Breiman. Random forests. Machine learning, 45(1):5–32, 2001.

[10] C. Chua and J. Wareham. Fighting internet auction fraud: An assessment and proposal. Computer, 37(10):31–37, 2004.

[11] R. Tibshirani. Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society. Series B (Methodological) ,58(1): 267-288, 1996.

[12] H. Ishwaran and J.Rao. Spike and Slab variable selection: Frequentist and Bayesian strategies. The Annals of Statistics,33(2): /730-733 ,2005.

[13] USA Today. How to avoid online auction fraud.
http://www.usatoday.com/tech/columnist/2002/05/07/yaukey.htm, 2002.

[14] C. Chua and J. Wareham. Fighting internet auctionfraud: An assessment and proposal. Computer, 37(10):31–37, 2004.

[15] P. Resnick, K. Kuwabara, R. Zeckhauser, and E. Friedman. Reputation systems. Communications of the ACM, 43(12):45–48, 2000.

[16] P. Resnick, R. Zeckhauser, J. Swanson, and K. Lockwood. The value of reputation on ebay: A controlled experiment. Experimental Economics, 9(2):79–101, 2006.

[17] D. Gregg and J. Scott. The role of reputation systems in reducing on-line auction fraud. International Journal of Electronic Commerce, 10(3):95–120, 2006.

[18] S. Pandit, D. Chau, S. Wang, and C. Faloutsos. Netprobe: a fast and scalable system for fraud detection in online auction networks. In Proceedings of the 16th international conference on World Wide Web, pages 201–210. ACM, 2007.

[19] P. McCullagh and J. Nelder. Generalized linear models. Chapman & Hall/CRC, 1989.

[20] C. Bliss. The calculation of the dosage-mortality curve. Annals of Applied Biology, 22(1):134–167, 1935.

[21] N. Cristianini and J. Shawe-Taylor. An introduction to support Vector Machines: and other kernel-based learning methods. Cambridge university press, 2006.

[22] J. Quinlan. Induction of decision trees. Machine learning, 1(1):81–106, 1986.

[23] A. Tikhonov. On the stability of inverse problems. In Dokl. Akad. Nauk SSSR, volume 39, pages 195–198,1943.

[24] R. Tibshirani. Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society. Series B (Methodological), 58(1):267–288, 1996.

[25] H. Ishwaran and J. Rao. Spike and slab variable selection: frequentist and bayesian strategies. The Annals of Statistics, 33(2):730–773, 2005.

[26] L. Wasserman. Bayesian model selection and model averaging. Journal of Mathematical Psychology, 44(1):92–107, 2000.

[27] C. Hans, A. Dobra, and M. West. Shotgun stochastic search for ¸Slarge pˇ T regression. Journal of the American Statistical Association, 102(478):507–516,2007.

[28] J. Friedman. Stochastic gradient boosting. Computational Statistics & Data Analysis, 38(4):367–378, 2002.

8. About the Authors

Krishnaveni Ampolu is currently pursuing her 2 Years M.Tech (CSE) in Department of Computer Science and Engineering at Avanthi's St.Theressa Institute of Engineering & Technology, Garividi, Vizianagaram District. Her area of interests includes Networks and Information Security.

P.Srinivasu Varma is currently working as Assistant Professor, in Department of Computer Science and Engineering at Avanthi's St.Theressa Institute of Engineering & Technology, Garividi, Vizianagaram District. His research interests include Networks Security & Information Security, Data Mining.

